Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Angela Brooks (anbrooks@ucsc.edu)
University of California, Santa Cruz https://orcid.org/0000-0002-7898-3073

Francisco Pardo-Palacios
Polytechnical University of Valencia https://orcid.org/0000-0002-3067-0166

Fairlie Reese
University of California, Irvine

Silvia Carbonell-Sala
Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology

Mark Diekhans
University of California, Santa Cruz https://orcid.org/0000-0002-0430-0989

Cindy Liang
University of California, Santa Cruz

Dingjie Wang
The Ohio State University

Brian Williams
California Institute of Technology https://orcid.org/0000-0003-3253-611X

Matthew Adams
University of California, Santa Cruz https://orcid.org/0000-0001-6793-6557

Amit Behera
University of California, Santa Cruz

Julien Lagarde
Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology

Haoran Li
The Ohio State University

Andrey Prjibelski
St. Petersburg State University

Gabriela Balderrama-Gutierrez
University of California, Irvine

Muhammed Hasan Çelik
University of California, Irvine

Maite De María

University of Florida

Nancy Denslow

University of Florida https://orcid.org/0000-0002-3946-3112

Natàlia Garcia-Reyero

US Army Engineer Research & Development Center

Stefan Goetz

Biobam Bioinformatics SL

Margaret Hunter

U.S. Geological Survey, Wetland, and Aquatic Research Center

Jane Loveland

European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus

Carlos Menor

Biobam Bioinformatics SL

David Moraga

University of Florida

Jonathan Mudge

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus

Hazuki Takahashi

RIKEN

Alison Tang

University of California, Santa Cruz

Ingrid Youngworth

Stanford University

Piero Carninci

RIKEN, Center for Integrative Medical Sciences https://orcid.org/0000-0001-7202-7243

Roderic Guigó

Pompeu Fabra University

Hagen Tilgner

Weill Cornell Medicine

Barbara Wold

California Institute of Technology https://orcid.org/0000-0003-3235-8130

Christopher Vollmers

University of California, Santa Cruz

Gloria Sheynkman

University of Virginia

Adam Frankish

European Bioinformatics Institute

Kin Fai Au

Department of Biomedical Informatics, The Ohio State University https://orcid.org/0000-0002-9222-4241

Ana Conesa

Spanish National Research Council (CSIC)

Ali Mortazavi

University of California, Irvine https://orcid.org/0000-0002-4259-6362

Analysis

Keywords: long-read RNA, sequence methods, LRGASP, transcriptome analyses

DOI: https://doi.org/10.21203/rs.3.rs-777702/v1

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Francisco J. Pardo-Palacios[1,29], Fairlie Reese[2,3,29], Sílvia Carbonell-Sala[4,29], Mark Diekhans[5,29], Cindy E. Liang[6,29], Dingjie Wang[7,29], Brian Williams[8,29], Matthew S. Adams[6], Amit K. Behera[9], Julien Lagarde[4], Haoran Li[7], Andrey D. Prjibelski[10], Gabriela Balderrama-Gutierrez[2,3], Muhammed Hasan Çelik[2,3], Maite De María[11,12], Nancy Denslow[13], Natàlia Garcia-Reyero[14], Stefan Goetz[15], Margaret E. Hunter[16], Jane E. Loveland[17], Carlos Menor[15], David Moraga[18], Jonathan M. Mudge[17], Hazuki Takahashi[19], Alison D. Tang[9], Ingrid Ashley. Youngworth[20], Piero Carninci[19,21], Roderic Guigó[4,22], Hagen U. Tilgner[23], Barbara J. Wold[8], Christopher Vollmers[9,30], Gloria M. Sheynkman[24,25,26,30], Adam Frankish[17,30], Kin Fai Au[7,30], Ana Conesa[27,28,30*], Ali Mortazavi[2,3,30*], Angela N. Brooks[5,9,30*]

[1]Department of Applied Statistics and Operational Research and Quality, Polytechnical University of Valencia, Valencia, Spain, [2]Developmental and Cell Biology, [3]Center for Complex Biological Systems, University of California, Irvine, Irvine, USA, [4]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain, [5]UC Santa Cruz Genomics Institute, [6]Molecular Cell and Developmental Biology, University of California, Santa Cruz, Santa Cruz, USA, [7]Department of Biomedical Informatics, The Ohio State University, Columbus, USA, [8]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, USA, [9]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, USA, [10]Center for Bioinformatics and Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia, [11]Department of Physiological Sciences, College of Veterinary Medicine, [12]Center for Environmental and Human Toxicology, [13]Department of Physiological Sciences, Center for Environmental and Human Toxicology, University of Florida, Gainesville, USA, [14]Environmental Laboratory, US Army Engineer Research & Development Center, Vicksburg, USA, [15]Biobam Bioinformatics SL, Valencia, Spain, [16]U.S. Geological Survey, Wetland, and Aquatic Research Center, Gainesville, USA, [17]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [18]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, USA, [19]Center for Integrative Medical Sciences, Laboratory for Transcriptome Technology, RIKEN, Yokohama, Japan, [20]Department of Genetics, Stanford University, Palo Alto, USA, [21]Human Technopole, Milano, Italy, [22]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain, [23]Brain and Mind Research Institute and Center for Neurogenetics, Weill Cornell Medicine, New York City, USA, [24]Department of Molecular Physiology and Biological Physics, [25]Center for Public Health Genomics, [26]UVA Cancer Center, University of Virginia, Charlottesville, USA, [27]Institute for Integrative Systems Biology, Spanish National Research Council (CSIC), Paterna, Spain, [28]Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, USA, [29]These authors contributed equally, [30]These authors jointly supervised the work, *correspondence: ana.conesa@csic.es, ali.mortazavi@uci.edu, anbrooks@ucsc.edu

Abstract

With increased usage of long-read sequencing technologies to perform transcriptome analyses, there becomes a greater need to evaluate different methodologies including library preparation, sequencing platform, and computational analysis tools. Here, we report the study design of a community effort called the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium, whose goals are characterizing the strengths and remaining challenges in using long-read approaches to identify and quantify the transcriptomes of both model and non-model organisms. The LRGASP organizers have generated cDNA and direct RNA datasets in human, mouse, and manatee samples using different protocols followed by sequencing on Illumina, Pacific Biosciences, and Oxford Nanopore Technologies platforms. Participants will use the provided data to submit predictions for three challenges: transcript isoform detection with a high-quality genome, transcript isoform quantification, and *de novo* transcript isoform identification. Evaluators from different institutions will determine which pipelines have the highest accuracy for a variety of metrics using benchmarks that include spike-in synthetic transcripts, simulated data, and a set of undisclosed, manually curated transcripts by GENCODE. We also describe plans for experimental validation of predictions that are platform-specific and computational tool-specific. We believe that a community effort to evaluate long-read RNA-seq methods will help move the field toward a better consensus on the best approaches to use for transcriptome analyses.

Introduction

There is a growing trend of using long-read RNA-seq (lrRNA-seq) data for transcript identification and quantification, primarily with Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) platforms[1–4]. Consequently, there is a need to evaluate these approaches for transcriptome analysis to compare the impact of different sequencing platforms, multiple sequencing library preparation methods, and computational analysis methods (Reviewed in [5–8]).

A previous effort by the RNA-Seq Genome Annotation Assessment Project (RGASP) Consortium[9,10] involved evaluating short-read Illumina RNA-seq for transcript identification and revealed limitations in recalling full-length transcript products due to the complexity of eukaryotic transcriptomes. Although lrRNA-seq should improve transcript reconstruction, at a fixed cost,

the reduced sequencing depth and higher error rates of long-read sequencing approaches may offset the improvements.

To evaluate long-read approaches for transcriptome analysis, we formed the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium modeled after the previous GASP[11], EGASP[12], and RGASP[9,10] efforts. For this project, we aim for an open community effort in order to be as transparent and inclusive as possible in evaluating technologies and computational methods (**Fig 1**).

The LRGASP Consortium will evaluate three fundamental aspects of transcriptome analysis. First, we will assess the reconstruction of full-length transcripts expressed in a given sample from a well-curated eukaryotic genome such as human and mouse. Second, we will evaluate the quantification of the abundance of each transcript. Finally, we will assess *de novo* reconstruction of full-length transcripts from samples without a high-quality genome, which would be beneficial for annotating genes in non-model organisms. These evaluations became the basis of the three challenges that comprise the LRGASP effort (**Box 1**).

Challenge 1: Transcript isoform detection with a high-quality genome
Goal: Identify which sequencing platform, library prep, and computational tool(s) combination gives the highest sensitivity and precision for transcript detection.

Challenge 2: Transcript isoform quantification
Goal: Identify which sequencing platform, library prep, and computational tool(s) combination gives the most accurate expression estimates.

Challenge 3: *De novo* transcript isoform identification
Goal: Identify which sequencing platform, library prep, and computational tool(s) combination gives the highest sensitivity and precision for transcript detection without a high-quality annotated genome.

Box 1: Overview of the LRGASP Challenges

The LRGASP Challenges will use data produced by the LRGASP Consortium Organizers (**Fig 1b, Table 1, Supplementary Table 1**). The samples for Challenges 1 and 2 consist of human

and mouse ENCODE biosamples with extensive chromatin-level functional data generated separately by the ENCODE Consortium. These include the human WTC-11 iPSC cell line and a mouse 129/Casteneus ES cell line for Challenge 1 and a mix of H1 and Definitive Endoderm derived from H1 (H1-DE) for Challenge 2. In addition, individual H1 and H1-DE samples are being sequenced on all platforms; however, those reads will not be released until after the end of the challenge. All samples were grown as biological triplicates with the RNA extracted at one site, spiked with 5'-capped Spike-In RNA Variants (Lexogen SIRV-Set 4), and distributed to all production groups. After sequencing, reads for human and mouse samples were deposited at the ENCODE Data Coordination Center (DCC) for community access, including but not limited to the challenges. A single replicate of manatee whole blood transcriptome was generated for Challenge 3. For each sample, we performed different cDNA preparation methods, including an early-access ONT cDNA kit (PCS110), ENCODE PacBio cDNA, R2C2[13] for increased sequence accuracy of ONT data, and CapTrap to enrich for 5'-capped RNAs. CapTrap is derived from the CAGE technique[14] and was adapated for lrRNA-seq (manuscript in preparation). We also performed direct RNA sequencing (dRNA) with ONT.

Sample	# of Reps	PacBio cDNA	ONT cDNA	ONT direct RNA	R2C2	CapTrap PacBio	CapTrap ONT	Illumina cDNA
Mouse 129/Cast ES cell line	**3**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Human WTC-11	**3**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Human H1 ES/Definitive Endoderm cell line mix	**3**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Human H1 ES cell line	*3*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Human H1 Definitive Endoderm cell line	*3*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Trichechus manatus peripheral blood mononuclear cells	**1**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	**Yes**

Table 1: Overview of LRGASP sequencing data. The H1 and H1 Definitive Endoderm samples are sequenced but are not available to participants until the close of challenges.

Participants may provide multiple submissions for each challenge (detailed in **Challenge submissions and timeline**) and in any or all challenges. We will compare solutions where only lrRNA-seq data was used and solutions that include additional publicly-available data. Depending on the challenge, they will submit either a GTF or quantification file, additional metadata, and a link to a repository (e.g., Github) where a working copy of the exact analysis pipeline used to generate their results can be downloaded. We expect to re-run analysis pipelines for well-performing submissions to help ensure reproducibility. The evaluation of the challenge will comprise both bioinformatics and experimental approaches. SQANTI3 (https://github.com/ConesaLab/SQANTI3) will be used to obtain transcript features and performance metrics that will be computed on the basis of SIRV-Set 4 spike-ins, simulated data, and a set of undisclosed, manually curated transcript models defined by GENCODE[15]. Human models will further be compared to histone modification ChIP-seq, open chromatin, CAGE, and poly(A)-seq results. Experimental validation will be performed on a select number of loci with either high agreement or disagreement between sequencing platforms or analysis pipelines. Evaluation scripts and experimental protocols will be publicly available in advance of submission deadlines (**Data and code availability**).

Methods

Additional details of all protocols for library preparation and sequencing can be found at the ENCODE DCC and is linked to each dataset produced by LRGASP (**Supplementary Table 1**).

Capping SIRVs

Exogenous synthetic RNA references (spike-ins) are widely used to calibrate measurements in RNA assays, but they lack the 7-Methylguanosine (m^7G) cap structure that most natural eukaryotic RNA transcripts bear at their 5' end. This characteristic makes commercial spike-in mixes unsuitable for library preparation protocols involving 5' cap enrichment steps. Therefore, we enzymatically added the appropriate m^7G structure to the SIRV standards used in this challenge. Specifically, the pp5'N structure present at the 5' end of spike-in sequence was used as a template for the Vaccinia capping enzyme (catalog num M2080S, New England BioLabs) to add the m^7G structure to SIRV-Set 4 (Iso Mix E0 / ERCC / Long SIRVs, catalog num 141.03,

Lexogen). A total of ten vials of SIRV-Set 4 (100 µl) were employed to perform the capping reaction (final total mass of 535 ng). The reaction was performed following the recommendations of the manufacturer's capping protocol with two minor changes: 3.5 µl of RNAse inhibitors (RNasin Plus RNase Inhibitor, catalog num N2611, Promega) were added to the capping reaction to avoid RNAse degradation, and the incubation time was extended from 30 minutes to two hours, following a recommendation from New England BioLabs technical support scientists. The final capping reaction was purified by using 1.8x AMPure RNA Clean XP beads (catalog num. A63987, Beckman Coulter) and resuspended in 100 µl of nuclease-free water.

Mouse and human RNA sample preparation

Prior to distribution of biosample total RNA aliquots to each of the participating labs, 110 µg of each biosample total RNA was spiked with Lexogen Long SIRV Set-4 quantification standards (catalog # 141.03) at approximately 3% of the estimated mRNA mass present (~1% of total RNA). The mass of capped SIRVs used was 29.5 ng and the mass of uncapped SIRVS used was 28.9 ng. In the case of direct RNA sequencing of one replicate of WTC-11 (ENCODE library accession ENCLB926JPE) and one replicate of mouse ES cells (ENCODE library accession ENCLB386NNT), only uncapped SIRV 4.0 were spiked in at approximately 3% of the estimated mass. Appropriate volumes of the spiked total RNA mixture to meet the input mass requirements for each library preparation method were then aliquoted separately, stored at -80 C, and shipped on dry ice to participating labs.

Manatee RNA sample preparation

Blood samples from Florida manatees were collected during health assessments by the U.S Geological Survey (USGS) Sirenia Project, the Florida Fish and Wildlife Conservation Commission (FWC), and the University of Florida under U.S. Fish and Wildlife Service (USFWS) permit # MA791721-5 in Crystal River (Citrus County, Florida, USA) and in Satellite Beach (Brevard County, Florida, USA) in December and January of 2018 and 2019 respectively. Samples were processed under the University of Florida USFWS permit #MA067116-2 following a protocol approved by the ethics committee (IACUC # 201609674 & IACUC # 201909674). Whole blood from minimally restrained Florida manatees were collected from the medial interosseous space between the ulna and radio from the pectoral flippers. Samples were drawn using Sodium Heparin 10-mL BD vacutainers (BD BioScience, New

Jersey, U.S.A). Blood samples were spun on-site and the plasma was aliquoted, stored in liquid nitrogen or ice, and transferred to -80 ºC once in the lab. The buffy coat (white blood cells) was flash-frozen in liquid nitrogen on-site and total RNA was extracted subsequently in the lab using STAT 60 (Tel-test Friendswood, TX) reagent. Approximately 350 µL of the frozen buffy coat was added to 1 ml of STAT 60 and vortexed for 30 seconds, 250 µL of chloroform was added and the tube was centrifuged 20,800 x g for 15 minutes at 4 ºC, to extract the RNA. This step was repeated and then RNA was precipitated from the supernatants overnight at -20ºC by the addition of 700 µL isopropanol with 1.5 µL of GlycoBlueTM (15 mg/mL) (Ambion, Invitrogen, Austin, TX) as a coprecipitant. Following centrifugation at 20,800 x g for 45 minutes, the pellet was washed with ethanol 70%, air-dried, and resuspended in 20 mL of RNA secure (Ambion, Austin, TX). A DNAse treatment was performed using Turbo DNA-freeTM kit (Ambion, Austin, TX). A total of nine good-quality RNA samples were selected to create an RNA pool. These samples included 6 females, one calf, one lactating female and one male and had RIN values from 8.0 to 8.8.

Manatee genome sample preparation

The genome of the Florida manatee Lorelei was sequenced using Nanopore and Pacbio. Lorelei is the same individual manatee for which an Illumina-based genome assembly was released by the Broad Institute in 2012 [16]. An EDTA, -80ºC whole blood sample aliquot was used. gDNA was extracted from 1400 µl of blood using the DNeasy kit (QIAGEN, MD, USA) following the companies' specifications for 100 µl aliquots of blood. Thawed blood was diluted 1:1 with RNA free Phosphate buffered saline 1x (Gibco, UK), 20 µl of proteinase K (QIAGEN, MD, USA), and 200 ul of AL lysis buffer (QIAGEN, MD, USA) and vortexed immediately. It was incubated at 56 °C for 10 minutes. Then, we added 200 µl of ethanol 96% and mixed it thoroughly. The mixture was added to the DNeasy mini spin-column and centrifuged at 6,000 x g for 1 minute. The column was washed with 500 µl of AW1 solution (QIAGEN, MD, USA) and centrifuged at 6,000 x g for 1 minute and followed with a wash with 500 µl AW2 (QIAGEN, MD, USA) and centrifuged 20,000 x g for 3 minutes. gDNA was eluted twice with 100 µl of AE buffer added to the center of the column, incubated for 1 minute, and centrifuged 6,000 x g for 1 minute. The first and second elution from the DNeasy mini spin-column were pooled and concentrated using a speed vacuum for 20 minutes in which each preparation was reduced from 200 to 50 µl. All gDNA tubes were pooled and the DNA was cleaned with AM Pure magnetic beads (Beckman Coulter-Life Sciences, IN, USA) at a ratio of 0.5:1, beads volume to

gDNA volume (50 µl of beads to 100 µl of gDNA). gDNA bound to the beads was washed twice with 1 ml of 70% ethanol. Ethanol traces were removed by quick spin to the bottom of the tube and removed with a pipette. Then, the beads were dried for 2 minutes and gDNA was eluded in 55 µl of EB buffer (QIAGEN, MD, USA) at 37 °C with 10 minutes of incubation. This process was repeated twice. Quantification of gDNA was performed with a QubitTM fluorometer (Thermo Fisher Scientific) and the quality of the gDNA was assessed using a Genomic Agilent TapeStation (Santa Clara, CA, USA). The final DNA quantity was 28.8 µg of DNA at a concentration of 267 ng/µl. The DNA Integrity Number (DIN) was 8.8 and the peak size was 54.5 kb.

cDNA preparation for Illumina and PacBio sequencing of human and mouse

PacBio cDNA synthesis was performed using a modified version of the Picelli protocol[17] with the Maxima H- reverse transcriptase. Total RNA was treated with exonuclease to remove transcripts without a cap. 2 µl of exonuclease-treated RNA were mixed with a priming reaction (RNAse inhibitor, dNTP's and water)was incubated at 72ºC for 3 minutes, then ramps down to 50ºC. While in the PCR block we added oligo dT (stock concentration 10 nM) and were incubated 3 min at 50ºC. We then added a first strand synthesis buffer (5x RT buffer, TSOligo, water) that had previously been incubated at 50ºC for one minute. The previous reaction was then incubated in the PCR block (Extension at 50ºC for 90 min, 85ºC for 5 min and held at 4ºC). To the same reaction we added a mix for amplification (2x reaction buffer, IS primers - 20 nM stock, water and SeqAmp polymerase). Then we ran a PCR program to amplify the cDNA (95ºC 1 min, 98ºC 15 sec, 65ºC 30 sec and 68ºC 13 min. The cycle repeats 10 times, which is followed by incubation at 72ºC for 10 min and holding at 4ºC. The amplified products were purified using SPRI beads and checked for quality in a bioanalyzer.

PacBio library preparation of human and mouse libraries

To build PacBio libraries, we started from 500 ng of polyA selected cDNA. The ends of the cDNA were repaired first in order for the cDNA molecule to be suitable for ligation of SMRTbell adapters. We added a damage repair reaction (DNA prep buffer, NAD and DNA damage repair) and then incubated at 37ºC for 30 min. Then End prep mix was added and incubated at 20ºC for 30 min and 65ºC 20 min. Ligation of the adapter at the ends of the cDNA was done by adding a ligation mix (pacbio adapters, ligation mix, ligation enhancer and ligation additive), then it was

incubated at 20ºC for 60 min. Final libraries were cleaned up using SPRI beads and we recorded the size and concentration of samples. Once the ligation step was done and the libraries passed the QC, a sequencing primer was annealed to the adapters in the UCI GHTF sequencing facility to allow for the binding of the polymerase during sequencing.

CapTrap preparation for PacBio and ONT sequencing of human and mouse

CapTrap is a technique developed by the Guigó laboratory (CRG, Barcelona, Spain) in collaboration with the group of Piero Carninci in RIKEN, Japan. The method enriches for full-length transcripts by selection of the 7-Methylguanosine (m7G) cap structure present at the 5' ends of RNA transcripts, followed by specific cap- and polyA- dependent linker ligations. The cDNA libraries generated using this method are compatible with long-read sequencing platforms (ONT or PacBio). The protocol starts with first strand synthesis (PrimeScript II Reverse Transcriptase, catalog num. 2690A, Takara) where 5 µg of total RNA polyA+ RNAs are fully reverse transcribed using a 16-mer anchored dT oligonucleotide. First strand synthesis was performed at 42 ºC for 60 minutes. Resulting products were purified with 1.8x AMPure RNA Clean XP beads (catalog num. A63987, Beckman Coulter). After the first-strand generation, the m7G cap structure at the 5' end of the transcripts is selectively captured using the CAP-trapper technique [14,18], which leads to the removal of uncapped RNAs. The diol group on the m^7G cap is oxidized with 1M NaOAc (pH 4.5) and NaIO4 (250 mM). Tris HCl (1M, pH 8.5) was added to stop the reaction and the whole reaction was purified with 1.8x AMPure RNA Clean XP beads. Aldehyde groups were biotinylated using a mixture containing NaOAc (1M, pH 6.0) and Biotin (Long Arm) Hydrazide (100 mM, catalog num. SP-1100, Vector Laboratories). The resulting mixture was then incubated for 30 minutes at 40ºC and purified with 1.8x AMPure RNA Clean XP beads. Single strand RNA was degraded by RNase ONE Ribonuclease (catalog num. M4261, Promega) for 30 minutes at 37ºC and purified with 1.8x AMPure RNA Clean XP beads. The m7G cap structure bound to biotin is then selected using M-270 streptavidin magnetic beads (catalog num. 65305, Thermo Fisher Scientific). M-270 streptavidin magnetic beads were equilibrated with CapTrap Lithium chloride/Tween 20 based binding buffer. Sample recovered after RNase ONE purification was bound to equilibrated M-270 streptavidin magnetic beads (incubation at 37ºC for 15 minutes), washed 3 times with CapTrap Tween20 based washing buffer and released by heat shock for 5 minutes at 95ºC and quickly cooled on ice. A second release was performed, and the supernatant was also collected and mixed with the eluate from the previous release. The released sample was treated with RNase H (60 U/µl, Ribonuclease H <RNase H>, catalog num. 2150,

Takara), RNase ONE (10 U/µl) and CapTrap release buffer (incubated at 37ºC for 30 minutes), purified with 1.8x AMPure XP beads (catalog num. A63881, Beckman Coulter) and concentrated by using a speed vac. After this cap specific selection, two double-stranded linkers, carrying a unique molecular identifier (UMI), are specifically ligated to the first strand cDNA [19]. Linker ligation (DNA Ligation Kit <Mighty Mix>, catalog num. 6023, Takara) was performed in two separate steps. First the 5' linker was ligated, purified twice, to completely eliminate the non-incorporated linkers, with 1.8x AMPure XP beads and concentrated by using a speed vac. Then the 3' linker was ligated, purified once with 1.8x AMPure XP beads and finally concentrated by using a speed vac. The double stranded linkers are converted into single strand by Shrimp Alkaline Phosphatase (1 U/µl SAP, catalog num. 78390, Affymetrix) and Uracil-Specific Excision Reagent (1 U/µl USER, catalog num. M5505L, NEB) treatment. This reaction was incubated for 30 minutes at 37ºC, 5 minutes at 95ºC and finally placed on ice. The sample was then purified with 1.8x AMPure XP beads. After this treatment, the two linkers which serve as priming sites for the polymerase (2x HiFi KAPA mix, catalog num. 7958927001-KK2601, Kapa), enable the synthesis of the full-length second strand. The mixture was incubated for 5 minutes at 95ºC, 5 minutes at 55ºC, 30 minutes at 72ºC and finally held at 4ºC until 1 µl Exonuclease I (20U/µl, catalog num. M0293S, NEB) was added to each sample. The sample was then incubated for 30 minutes at 37ºC and afterwards, purified twice with 1.8x and 1.4x (respectively) AMPure XP beads and finally concentrated in a speed vac. The resulting cDNA is amplified (TaKaRa LA Taq, catalog num. RR002M, Takara) via long and accurate PCR (LA PCR) protocol. In order to avoid PCR duplicates, each sample was split in two PCR independent reactions and amplified 16 cycles with 15 seconds at 55ºC for annealing, and 8 minutes at 65ºC for extension. The 2 PCR replicates were merged and purified with 1x AMPure XP beads. Samples were quantified with Qubit (Qubit 4 Fluorometer, Thermo Fisher Scientific) and quality-checked with BioAnalyzer (Agilent 2100 Bioanalyzer, Agilent Technologies).

CapTrap MinION cDNA sequencing was performed with 500 ng of cDNA sample coming from CapTrap cDNA protocol and strictly following the SQK-LSK109 adapter ligation protocol (ONT). The cDNA sequencing on MinION platform was performed using ONT R9.4 flow cells and the standard MiniKNOW protocol.

PacBio Sequel II sequencing was performed using 500 ng of CapTrap samples following the SMRTbell™ Express Template Prep Kit 2.0 protocol.

R2C2 preparation for ONT sequencing of human and mouse

For each biological replicate, two libraries were created, a regular (non-size selected), and a size selected library of cDNA over 2 kb in length to achieve higher coverage of longer transcripts. For each RNA sample, 400 ng was used to generate full-length single stranded cDNA using an indexed oligo(dT) primer and a template switching oligo (TSO). PCR was used to generate the second strand and amplify the library. The cDNA was then isolated by SPRI bead clean up. For the size selected libraries, cDNA was run on a 1% low melt agarose gel. A smear in the range of 2–10 kb was excised from the gel and digested with beta-agarase followed by SPRI bead clean up. At this point, indexed cDNA from each biological replicate was pooled together equally. cDNA was circularized using a short DNA splint with sequence complementary to the cDNA ends by Gibson Assembly (NEBuilder, NEB) with a 1:1 cDNA:splint ratio (100 ng each). After Gibson assembly, a linear digestion (ExoI, ExoIII, and Lambda Exonuclease) was performed to eliminate non-circularized DNA. The circular Gibson assembly product was cleaned up using SPRI beads. The circularized library was used as template for rolling circle amplification (RCA) using Phi29 polymerase and random hexamer primers. Following the RCA reaction, T7 endonuclease was used to debranch the DNA product. A DNA clean and concentrator column was used to purify the DNA. Purified RCA product was size-selected using a 1% low melt agarose gel. The main band just over the 10 kb marker was excised from the gel and digested with beta-agarase followed by SPRI bead clean up. The cleaned and size selected RCA product was sequenced using the ONT 1D Genomic DNA by Ligation sample prep kit (SQK-LSK109) and MinION flow cells (R9.4.1) following the manufacturer's protocol. Flow cells were nuclease flushed and reloaded with additional library following ONT Nuclease Flush protocol.

cDNA preparation for ONT sequencing of human and mouse

Library preparation was done from total RNA (200ng) using SQK-PCS110 kit from ONT for PCR-cDNA sequencing. Briefly, cDNA RT adapters were annealed and ligated to full length RNAs using NEBNext® Quick Ligation Reaction Buffer (NEB B6058) and T4 DNA Ligase (NEB M0202). Bead clean up was done using Agencourt RNAClean XP beads. Purified RNA with CRTA top strand, RT primers, and dNTPs (NEB N0447) were incubated at RT for 15 mins to generate primer-annealed RNA. Reverse transcription and strand-switching was performed with Maxima H Minus RT enzyme in presence of strand-switching primers at 42^0C for 90 mins followed by heat inactivation at 85^0C for 5 mins. Reverse transcribed samples were PCR

amplified using cDNA primers and LongAmp Hot Start Master Mix (NEB, M0533S). Samples were treated with NEB exonuclease I (NEB, M0293) for 15 mins at 37^0C to degrade linear single-stranded DNA, followed by enzyme inactivation at 80^0C for 15 mins. Samples were purified with Agencourt AMPure XP beads. Elution was done with 12 ul of elution buffer. 1ul of libraries was electrophoresed on TapeStation screentapes to assess size distribution, quantity and quality of library. FLO-MIN106D flow cells were primed with EXP-FLP002 kit reagents followed by loading of PCR-cDNA library mixed with rapid adapter F (along with sequencing buffer and loading beads). Sequencing of the library was performed without any size selection using MinION Mk1B devices and MinKNOW software interface.

dRNA preparation for ONT sequencing of human and mouse

dRNA libraries were prepared from 75ug total RNA. RNA samples were poly-A selected using the NEXTFLEX poly-A kit. Purified mRNA was eluted in 12uL NF H2O. Library preparation was performed on purified mRNA using the SQK-RNA002 kit. Direct RNA RT adapters were annealed and ligated to full-length mRNA using T4 DNA Ligase, NEBNext Quick Ligation Reaction Buffer, and Nanopore's RNA CS. Adapter-ligated mRNA was incubated with dNTPs, 5x first-strand buffer, nuclease-free water, SuperScript IV, and 0.1M DTT to create a cDNA-RNA hybrid. This reverse-transcription (RT) step is recommended by Nanopore to reduce secondary structure formation of the mRNA as it is being sequenced. RTed RNA was purified using RNAClean XP beads. Nanopore adapters were ligated onto the RTed RNA using NEBNext Quick Ligation Reaction Buffer and T4 DNA Ligase. Following RNAClean XP bead cleanup, the libraries were eluted in 21uL of Nanopore's Elution Buffer. 1 uL of each library was quantified on the TapeStation to ensure nucleic acid concentration was at minimum ~200ng. Libraries were loaded into MinION flow cells using the EXP-FLP002 Flow Cell Priming Kit. Libraries were sequenced for 72 hour runs.

Manatee ONT genome sequencing

2 µg of genomic DNA in a total volume of 100 µl was fragmented with the g-Tube fragmentation method (Covaris, Woburn, MA, USA) by using centrifugation at 6,000x g for 1 min. The large DNA fragments were enriched by using 0.85x volume of Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) in the purification procedure. The enriched DNA fragments were subjected to library preparation with Nanopore Genomic DNA Ligation Sequencing Kit

(Oxford Nanopore Technologies, Oxford, UK) following the manufacture's protocol. A total of 700 ng of final library product was loaded on a flow cell and sequenced with a Nanopore GridION sequencer (Oxford Nanopore Technologies, Oxford, UK) for a 72-hr run. A total of 5 flow-cell runs were conducted for this project.

Manatee cDNA Pacbio library preparation and sequencing

Approximately 280 ng of total pooled RNA were processed according to a modified IsoSeq protocol. The sample was spiked-in with the uncapped E2 RNA variant control mix (SIRVs, Lexogen, Cat # 025.03) at a 2.83% mass proportion relative to the total RNA. The resulting mixture was subjected to a globin removal step using the QIAseq FastSelectTM- HRM Globin removal reagent (cat # 334376). This kit was designed for globin removal from human, mouse, and rat tissues and was found to perform with various degrees of efficiency on blood from a wide variety of samples of mammalian origin. Globin removal was performed as recommended in the QIAseq FastSelectTM- -rRNA HRM -Globin Handbook (Oct 2019) in the NEBNext Ultra II section, except that the high-temperature fragmentation step was omitted. The globin removal reaction (9 µl) contained: 280 ng sample (RNA plus 2.83% SIRVs), QIAseq FastSelect globin removal reagent, 2 µl NEBNext Single Cell RT Primer Mix (NEB #6421), and 2.25 µl of NEBNext Single Cell RT buffer (4x). This mixture was prepared in a 0.2 ml PCR tube and subjected to a stepwise series of 2 min incubations each of 75°C, 70°C, 65°C, 60°C, 55°C, 37°C and 25°C. At this point, the sample was snap-cooled by transferring to a pre-chilled freezer block until ready for the RT and amplification steps. From this point on, cDNA synthesis was done as described in the "Protocol for Low Input RNA: cDNA Synthesis and Amplification" (NEB #E6421) starting on section 2.3. More specifically, the template "RT and Template Switching" reaction consisted of 9 µl of globin-removed RNA, 2.75 µl NEBNext Single Cell RT Buffer (4x), 1 µl of NEBNext Template Switching Oligo, 2 µl of NEBNext Single Cell RT Enzyme Mix and enough water to bring the total to 20 µl. The reaction was incubated in a thermocycler for 90 min at 42 °C and 10 min at 72 °C. The cDNA products were split into four aliquots for PCR amplification (100 µl) reactions containing 2 µl NEBNext Single Cell cDNA PCR Primer, 0.5 µl 10X NEBNext Cell Lysis Buffer, 50 µl NEBNext Single Cell cDNA PCR Master Mix, 5 µl RT and Template Switching reaction and water. Amplified cDNA was purified by AMPure, one round at 0.8 to 1.0 beads to sample ratio and one round at 0.65:1.0 ratio. The yield of amplified cDNA by this modified protocol (300-400 ng) was about 10-fold lower than the standard protocol (i.e., without globin-removal). The average cDNA size was ~1400 bp. When increased amounts of cDNA were desired the cDNA was amplified by 5 additional PCR cycles.

Two preps obtained with the above described protocol were pooled together and 500 ng were loaded on an electrophoretic lateral fractionation system (ELF, SageScience). Fragments above 2.5 kb were collected, re-amplified (10 cycles), and re-pooled equimolarly with non-size-selected cDNA fragments. This re-pooled cDNA prep is referred to as “enriched cDNA_>2.5kb”. Both non_enriched cDNA and enriched cDNA_>2.5kb cDNA were used for SMRT bell library was constructed starting with 1 µg of cDNA as described (PacBio IsoSeq protocol 101-070-200 Version 06, September 2018). Briefly, SMRTbell adaptors (Iso-SeqTM) were added using reagents from the PacBio SMRTbell Template Prep Kit 1.0-SPv3 starting with either 200 ng (for enriched cDNA >2.5kb) or 700 ng (for non enriched cDNA). The main steps included: DNA Damage Repair, End Repair, Blunt-end ligation of SMRT bell adaptors, and ExoIII/ExoVII treatment. This procedure resulted in ~25-30% yield. Finally, libraries were eluted in 15 ul of 10 nM Tris HCl, pH 8.0. Library fragment size was estimated by the Agilent TapeStation (genomic DNA tapes), and this data was used for calculating molar concentrations.
The enriched cDNA >2.5 kb library was diffusion-loaded on a single SEQUEL SMRT cell (University of Florida, ICBR-NGS core lab) at loading concentration was 10 pM, using 4-hr pre-extension, 20 hr movies and v3 chemistry reagents (for binding and sequencing). All other steps for sequencing were done according to the recommended protocol by the PacBio SMRT Link Sample Setup and Run Design modules (SMRT Link 6.0).

The non enriched cDNA library loaded on three Sequel II SMRT cells at University of California, Irvine.

Manatee cDNA Nanopore library preparation and sequencing

One hundred and fifty nanograms of total pooled RNA were processed according to a modified ONT cDNA-PCR Sequencing protocol (cDNA-PCR-PCS109, version PCS_9085 v109 revJ Aug 14, 2019). Spike-in and globin depletion treatment was conducted as described for Pacbio library preparation. In this case, the globin removal reaction (11 ul) contained: sample (RNA plus SIRVs), globin removal reagent, 1 mM dNTP, 0.2 µM VPN primer from the Nanopore cDNA synthesis protocol (i.e., in place of random primers), and 1X RT buffer (ThermoFisher). This mixture was prepared in a 0.2 ml PCR tube and submitted to a stepwise series of 2 min incubation for each of 75 °C, 70 °C, 65 °C, 60 °C, 55 °C, 37 °C and 25 °C. At this point, the sample was snap-cooled by transferring to a pre-chilled freezer block until ready for the RT and amplification steps. From this point on, cDNA synthesis was done as described in the cDNA-PCR Sequencing (SQK-PCS109) Nanopore manual starting on page 9 (Version: PCS_90985_v109_revJ_14Aug2019). A single globin removal and cDNA synthesis reaction was split into four PCR reactions for amplification.

This process resulted in approximately 2 micrograms of “full-length” cDNA with an average size of ~1800 bp. One size-selected library was constructed by loading 1500 ng of this cDNA on an electrophoretic lateral fractionation system (ELF, SageScience), collecting.5 kb were collected, re-ampliying (6 cycles) and re-pooling with non-size-selected cDNA fragments. Adaptor ligation and sequencing were performed according to the cDNA-PCR Sequencing (SQK-PCS109) Nanopore manual. Between 120-140 fmol of cDNA was loaded on a FLO-MIN106D (R9.4 SpotON) flow cell for sequencing on the minION device. Two runs were done on non-size-selected manatee cDNA, while only one run was done on the cDNA that had been enriched with >2.5 kb fragments. Sequencing runs were allowed to proceed for 48 hours.

Long-read data processing

Basecalling of ONT data from human, mouse and manatee was performed with Guppy 4.2.2 and hac 9.4.1 config file, with default parameters, except: --qscore_filtering --min_qscore 7 (these non-default parameters were used in all ONT cDNA runs except for R2C2 datasets). Direct RNA basecalling was also performed with Guppy 4.4.2 with the following configurations: --qscore_filtering yes --min_qscore 7 --reverse_sequence yes
--u_substitution yes

PacBio full-length non-chimeric (FLNC) reads were generated with CCS 4.2.0 (parameters: --noPolish --minLength=10 --minPasses=3 --min-rq=0.9 --min-snr=2.5), Lima 1.11.0 (parameters: FASTA with the appropriate adapters --isoseq --min-score 0 --min-end-score 0 --min-signal-increase 10 --min-score-lead 0), and Refine 3.3.0 (parameters: --min-polya-length 20 --require-polya).

Consensus R2C2 reads were generated with C3POa v1.0.0 (https://github.com/rvolden/C3POa/tree/gonk) with default options

Sequence data are provided in FASTQ format. For PacBio data, subreads are provided in unaligned BAM format and for R2C2 data, subreads are provided in FASTQ (**Supplementary Table 1).**

Reference genome and annotations

For submissions of transcript models and quantification, transcript annotations and genome models corresponding to GENCODE human v38 and mouse M27 will be used. Submissions of challenge predictions are expected to end in Fall 2021, prior to the release of GENCODE human v39 and mouse M28. The newly released GENCODE annotations will, therefore, be used for the evaluations. GRCh38 is the reference genome sequence for human and GRCm39 for mouse, GENCODE annotations are based on these genomes. Please note that GENCODE M25 and earlier annotation releases are based on GRCm38.

Simulated data

Simulating RNA reads simply from the reference transcriptome would only allow the assessment reconstruction of known transcript models. Thus, we extended both human and mouse annotations with artificial novel transcripts. To obtain those, we mapped reference transcripts of an undisclosed mammalian organism to the human and mouse genomes and converted the alignments into transcript models using SQANTI[20]. We then arbitrarily selected isoforms of known genes that have only canonical splice sites (GT-AG, GC-AG and AT-AC) and merged them into human and mouse GENCODE Basic annotations.

To generate realistic isoform expression profiles we selected undisclosed human and mouse long read datasets and quantified them simply by mapping to the reference transcripts with minimap2 v2.17 (Li, 2018). Artificial novel isoforms were assigned arbitrary expression values. Generated expression profile was further used for simulating short and long reads.

To simulate reads produced by different sequencing platforms we used existing simulation methods. Illumina 2x150bp read pairs were generated with the RSEM simulator[21] using an error model obtained from real RNA-Seq data[22] (accession number ERR1474891). ONT reads were simulated with NanoSim[23] using pre-trained cDNA and dRNA models available in the package with average error rate of 15.9% (4.8% substitutions, 6.0% deletions, 5.1% insertions) and 11.2% (2.8% substitutions, 5.9% deletions, 2.5% insertions) respectively. PacBio CCS reads were obtained with IsoSeqSim (https://github.com/yunhaowang/IsoSeqSim), which truncates input reference transcript sequences and uniformly inserts errors according to given probabilities. We used Sequel II truncation probabilities provided along with the package. Error rate was estimated using real PacBio cDNA CCS reads obtained in this work as 1.6% (0.4%

substitutions, 0.6% deletions, 0.6% insertions). Additionally, polyA tails were attached to the 3' end of reference transcript sequences prior to running the simulation.

We simulated two datasets containing reads from all 3 platforms listed above but with slightly different properties. Human datasets were simulated with 100 million Illumina read pairs, 30 million ONT cDNA and 10 million PacBio reads. Mouse datasets also contained 100 million Illumina read pairs, but equal amounts of PacBio CCS and ONT dRNA reads were generated (20 million sequences each).

To allow users to simulate their own data, the methods described above are implemented as simple command-line scripts which are available at https://github.com/LRGASP/lrgasp-simulation/.

CAGE data of WTC-11 samples for validation of transcript 5' ends

CAGE data from WTC-11 samples are being produced for validation of transcript 5' ends; therefore, will not be released until the close of the challenge submissions. CAGE data will be obtained from two RNA biological replicates of WTC-11, from the same exact RNA used for long-read sequencing.

The 15 µg of WTC-11 RNAs from each biological replicate, ENCODE BioSample Accession #ENCBS944CBA and #ENCBS474NOC, were used for the single strand (ss)CAGE library preparation described in the published protocol[24]. Briefly, the 15 µg RNAs were aliquoted to 5 µg in three tubes and reverse transcribed to cDNAs with random primers, and the RNA-cDNA hybrids were cap-trapped by the streptavidin beads. The single strand cDNAs were released from the beads and ligated to the Illumina adaptors with an index. 1080 amols of the cap-trapped single strand cDNAs from each biological replicate were sequenced by Illumina HiSeq Rapid SBS Kits v2 (SR, 150 cycles, 1 lane for each), producing approximately 40 million reads per sample.

QuantSeq of human and mouse samples for validation of transcript 3' ends

QuantSeq data (3' end sequencing) from challenge 1 and 2 samples are being produced for validation of 3' ends; therefore, this data will not be released until the close of the challenge

submissions. Data will be obtained from two RNA biological replicates of WTC-11, from the same exact RNA used for long-read sequencing.

Full-length transcript validation with NRCeq

Depending on sample availability, we may further sequence the WTC-11 cell line using NRCeq, a method for direct RNA sequencing which can distinguish full length reads (i.e., dRNA reads containing both the 5' cap and polyA tail)[25]. NRCeq uses an oligomer adaptation approach to ligate an adapter specifically to 5′ m_7G capped RNAs and performs polyA-selected direct RNA sequencing. NRCeq would provide additional data to validate start-to-end RNA transcript sequence without RT-PCR artifacts[26]. Because the technique requires approximately 2.5 ug of poly(A)-selected RNA, the sequencing will be performed on two independent biological replicates of WTC-11 that were not from the original cell batch from which long-read sequencing was performed.

GENCODE benchmarks and computational evaluation

Full manual annotation will be undertaken on 50 selected loci on both the human and mouse reference genomes. Transcript models will only be annotated during this exercise based on their support from long transcriptomic datasets generated by the consortium specifically for LRGASP. That is, no transcript annotation will be based on transcriptomic data from externally produced datasets, although annotators will use any publicly available orthogonal data to aid interpretation of aligned consortium data. For example, Fantom 5 CAGE datasets will be used to help identify transcription start sites and transcript 5' ends and RNA-seq-supported introns derived from high throughput reanalysis pipelines such as Recount will be used to support putative introns identified in the alignments of long transcriptomic data.

Manual annotation will be performed according to the guidelines of the HAVANA (Human And Vertebrate Analysis aNd Annotation) group[15,27]. Transcriptomic data will be aligned to the human and mouse reference genome using appropriate methods. We will test the benefits of aligning the transcriptomic data using multiple methods to reduce the impact of alignment errors and artefacts.

Annotators will also take advantage of local alignment tools integrated into annotation software to give further alternative views of alignments and improve annotation accuracy. Transcript models will be manually extrapolated from the alignments by annotators using the otter annotation interface [28]. Alignments will be navigated using the Blixem alignment viewer [29,30] and where required visual inspection of the dot-plot output from the Dotter tool[31] will be used to resolve any alignment with the genomic sequence that was unclear or absent from Blixem. Short alignments (<15 bases) that cannot be visualized using Dotter will be detected using Zmap DNA Search[31] (essentially a pattern matching tool). The construction of exon-intron boundaries will require the presence of canonical splice sites (defined as GT-AG, GC-AG and AT-AC) and any deviations from this rule will be given clear explanatory tags (for example non-canonical splice site supported by evolutionary conservation). All non-redundant splicing transcripts at an individual locus will be used to build transcript models, and all alternatively spliced transcripts will be assigned an individual biotype based on their putative functional potential. Once the correct transcript structure has been ascertained the protein-coding potential of the transcript will be determined on the basis of its context within the locus, similarity to known protein sequences, the sequences of orthologous and paralogous proteins, candidate coding regions (CCRs) identified by PhyloCSF, evidence of translation from mass spectrometry and Ribo-seq data, the presence of Pfam functional domains, the presence of possible alternative ORFs, the presence of retained intronic sequence and the likely susceptibility of the transcript to nonsense-mediated mRNA decay (NMD). Although the annotation of transcript functional biotype and CDS is not required of submitters, it will be added to transcripts as a matter of routine manual annotation and may be used to investigate the detection or non-detection of groups of transcripts by submitters. Where necessary, annotations will be checked by a second annotator to ensure completeness and consistency of annotation between the genes annotated for LRGASP and the remainder of the Ensembl/GENCODE geneset.

Computational evaluation of transcript isoform detection and quantification

Challenge 1 Evaluation: Transcript isoform detection

Four sets of transcripts will be used for evaluation of transcript calls made on human and mouse lrRNA-seq data

1. Lexogen SIRV-Set 4 (SIRV-Set 3 plus 15 new long SIRVs with sizes ranging from 4 to 12 kb)

2. Comprehensive GENCODE annotation: human v39, mouse vM28. GENCODE human v28 and vM27 are available at the time of the LRGASP data release and new versions of GENCODE will be released after the close of LRGASP submissions.
3. A set of transcripts from a subset of undisclosed genes which will be manually curated by GENCODE. These transcripts will thus be considered high-quality models derived from LRGASP data
4. Simulated data for both Nanopore (Nanosim) and PacBio (Iso-SeqSim) reads

The rationale for including these different types of transcript data is that each set creates a different evaluation opportunity, but also has its particular limitations. For example, SIRVs and simulated data provide a clear ground truth that allows the calculation of standard performance metrics such as sensitivity, precision or false discovery rate. Evaluation of SIRVs can identify potential limitations of both library preparation as well as sequencing, but the SIRVs themselves represent a dataset of limited complexity. Higher complexity can be generated when simulating long reads based on actual sample data. However, read simulation algorithms only capture some potential biases of the sequencing technologies (e.g., error profiles) and not of the library preparation protocols. In any case, both types of data approximate, but do not fully recapitulate real-world datasets. Evaluation against the GENCODE annotation[15] represents this real dataset scenario, although in this case the ground truth is not entirely known. This limitation will be partially mitigated by the identification of a subset of GENCODE transcript models that will be revised and deemed as high-confidence by GENCODE curators, and by follow-up experimental validation for a small set of transcripts using semi-quantitative RT-PCR and quantitative PCR (qPCR) approaches. In this way, although an exhaustive validation of the real data is not possible, estimates of the methods' performances can be inferred. By putting together evaluation results obtained with all these different benchmarking datasets, insights will be gained on the performance of the library preparation, sequencing and analysis approaches both in absolute and in relative terms.

The evaluation of the transcript models will be guided by the use of SQANTI categories[20] (**Fig 2a**), implemented in the SQANTI3 software (https://github.com/ConesaLab/SQANTI3), and will incorporate additional definitions and performance metrics to provide a comprehensive framework for transcript model assessment (**Table 2**). The evaluation considers the accuracy of the transcript models both at splice junctions and at 3'/ 5' transcript ends. It will take into account external sources of evidence such as CAGE data, polyA annotation and support by

Illumina reads (**Fig 2b**). A number of novel transcripts detected by all or most pipelines, as well as pipeline-, platform-, or library- preparation specific transcripts will be selected for experimental validation and manual review by the GENCODE project. The evaluation script is provided to participants (**Data and code availability**).

Table 2: Transcript Classifications and Definitions used by the LRGASP computational evaluation

Classification	**Description**
Full Splice Match (FSM)	Transcripts matching a reference transcript at all splice junctions
Incomplete Splice Match (ISM)	Transcripts matching consecutive, but not all, splice junctions of the reference transcripts
Novel in Catalog (NIC)	Transcripts containing new combinations of 1) already annotated splice junctions, 2) novel splice junctions formed from already annotated donors and acceptors, or 3) unannotated intron retention
Novel Not in Catalog (NNC)	Transcripts using novel donors and/or acceptors
Reference Match (RM)	FSM transcript with 5´ and 3´ends within 50 nts of the transcription start site (TSS)/transcription termination site (TTS) annotation
_3´_polyA_supported	Transcript with polyA signal sequence support or short-read 3' end sequencing (e.g. QuantSeq) support at the 3´end
_5´_CAGE_supported	Transcript with CAGE support at the 5´end
_3´_reference_supported	Transcript with 3´end within 50 nts from a reference transcript TTS
_5´_reference_supported	Transcript with 5´end within 50 nts from a

	reference transcript TSS
Supported Reference Transcript Model (SRTM)	FSM/ISM transcript with 5´ end within 50 nts of the TSS or has CAGE support AND 3´ end within 50 nts of the TTS or has polyA signal sequence support or short-read 3' end sequencing support
Supported Novel Transcript Model (SNTM)	NIC/NNC transcript with 5´ end within 50 nts of the TSS or CAGE support AND 3´ end within 50 nts of the TTS or has polyA signal sequence support or short-read 3' end sequencing support AND Illumina read support at novel junctions
% Long Read Coverage (%LRC)	Fraction of the transcript model sequence length mapped by one or more long reads
Redundancy	# LR transcript models / reference model
Longest Junction Chain ISM NIC / NNC	 # junctions in ISM / # junctions reference # reference junctions / # junctions in NIC/NNC
Intron retention (IR) level	Number of IR within the NIC category
Illumina Splice Junction (SJ) Support	% SJ in transcript model with Illumina support
Full Illumina Splice Junction Support	% transcripts in category with all SJ supported
% Novel Junctions	# of new junctions / total # junctions
% Non-canonical junctions	# of non-canonical junctions / total # junctions
% Non-canonical transcripts	% transcripts with at least one non-canonical junction
Intra-priming	Evidence of intra-priming (described in [20])
RT-switching	Evidence of RT-switching (described in [20])

629

Given these definitions, evaluation metrics are specified for each type of data.

SIRVs

In order to evaluate SIRVs, we will extract from each submission all transcript models that associate to SIRV sequences after SQANTI3 analysis. This not only includes FSM and ISM isoforms of SIRVs, but also NIC, NNC, antisense and fusion transcripts mapping to SIRV loci. The metrics for SIRV evaluation are defined as follows.

Table 3: Metrics and definitions for evaluation against SIRVs

SIRV_transcripts	Transcripts mapping to a SIRV chromosome
Reference SIRV (rSIRV)	Ground truth SIRV model
True Positive detections (TP)	rSIRVs identified as RM
Partial True Positive detections (PTP)	rSIRVs identified as ISM or FSM_non_RM
False Negative (FN)	rSIRVs without FSM or ISM
False Positive (FP)	NIC + NNC + antisense + fusion SIRV_transcripts
Sensitivity	TP/rSIRVs
Precision	RM/SIRV_transcripts
Non_redundant Precision	TP/SIRV_transcripts
Positive Detection Rate	unique(TP+PTP)/rSIRVs
False Discovery Rate	(SIRV_transcripts - RM)/SIRV_transcripts
Redundancy	(FSM + ISM)/unique(TP+PTP)

Simulated Data

The simulated data contains both transcript models based on the current GENCODE annotation and a number of simulated novel transcripts that will result in true NIC and NNC annotations.

Transcript models generated from simulated data will be analysed by SQANTI3 providing a GTF file that includes all simulated transcripts (GENCODE and novel) and excludes all transcripts for which reads were not simulated. The evaluation metrics for simulated data are defined as follows:

Table 4: Metrics and definitions for evaluation against simulated data

P	All simulated transcripts
True Positive (TP) TP_ref TP_novel	RM RM to GENCODE models RM to simulated novel transcript models
Partial True Positive (PTP) PTP_ref PTP_novel	ISM or FSM_non_RM ISM or FSM_non_RM of GENCODE models ISM or FSM_non_RM of simulated novel models
False Negative (FN) FN_ref FN_novel	Simulated transcripts without RM or PTP calls Simulated GENCODE models without RM or PTP calls Simulated novel models without RM or PTP calls
False Positive (FP)	NIC + NNC + antisense + fusion
Sensitivity Sens_ref Sens_novel	 TP_ref/P(GENCODE) TP_novel/P(Simulated novel)
Precision	TP/(TP+PTP+FP)
Positive Detection Rate	(TP+PTP)/P
False Discovery Rate	(FP+PTP)/(TP+PTP+FP)
Redundancy	# FSM and ISM per simulated transcript model

Comprehensive GENCODE annotation

Submitted transcript models will be analyzed with SQANTI3 using the newly released GENCODE annotation and different metrics will be obtained for FSM, ISM, NIC, NNC and Other models according to the scheme depicted below. Transcripts from new genes included in the latest annotation release will be catalogued as “Intergenic” initially, but considered FSM, ISM, NIC or NNC with an updated GENCODE annotation. This will allow evaluation of gene and transcript discovery on unannotated regions.

Table 5: Metrics for evaluation against GENCODE annotation

Metric	**FSM**	**ISM**	**NIC**	**NNC**	**Others**
Count	X	X	X	X	X
Reference Match (RM)	X				
_3´_polyA_supported	X	X	X	X	
_5´_CAGE_supported	X	X	X	X	
_3´_reference_supported	X	X	X	X	
_5´_reference_supported	X	X	X	X	
Supported Reference Transcript Model (SRTM)	X	X			
Supported Novel Transcript Model (SNTM)			X	X	
Distance (nts) to TSS/TTS of matched transcript	X	X			
Redundancy	X	X			
% Long Read Coverage (%LRC)	X				
Longest Junction Chain		X	X	X	
Intron retention level		X	X		
Illumina Splice Junction Support	X	X	X	X	X
Full Illumina Splice Junction Support	X	X	X	X	X
% Novel Junctions			X	X	

% Non-canonical junctions	X	X	X	X	X
% Transcripts with non-canonical junctions	X	X	X	X	X
Intra-priming	X	X	X	X	X
RT-switching	X	X	X	X	X
Number of exons	X	X	X	X	X

High-confidence transcripts derived from LRGASP data (Positives P are the set of all high-confidence transcripts)

Finally, a set of manually curated transcript models will be used to estimate sensitivity on real data. Metrics that will be applied in this transcript set are: TP, PTP, FN, Sensitivity, Positive Detection Rate, Redundancy and %LRC.

Challenge 2 Evaluation: Transcript isoform quantification

We will evaluate transcript isoform quantification performance with both simulated and real sequencing data, which includes SIRV-Set 4. While the ground truth is known for the simulated data and SIRV-Set4, we will experimentally quantify the abundances of transcript isoforms from select loci (genes) within the LRGASP samples. Specifically, we will interrogate the presence of specific transcript isoforms using qPCR measurements of isoform-specific regions, and will obtain such data using an aliquot of the exact same RNA which was used to generate the LRGASP datasets (human and mouse).

Evaluation metrics

We evaluate the quantification performance for different data scenarios (**Figure 3**):

1) Single sample data when the ground truth is available

2) Multiple replicates under two different conditions when the ground truth is available

3) Multiple replicates when ground truth is not available

The participants of the Challenge 2 can run these evaluations via submitting their quantification results at the website https://lrrna-seq-quantification.org/ that generates an interactive report in the html and PDF formats (See **Data and code availability**).

Single sample data (ground truth is available)

We can evaluate how close the estimations and the ground truth values are by four metrics as follows.

Denote $\hat{\Theta} = \left(\hat{\theta}_1, \cdots, \hat{\theta}_I\right)^T$ and $\Theta = (\theta_1, \cdots, \theta_I)^T$ as the estimation and ground truth of the abundance of I transcript isoforms in a sample, respectively. Then, four metrics can be calculated by the following formulas.

•Spearman Correlation Coefficient (*SCC*)

SCC evaluates the monotonic relationship between the estimation and the ground truth, which is based on the rank for transcript isoform abundance (**Supplementary Fig. S1**). It is calculated by

$$SCC_{\Theta,\hat{\Theta}} = \frac{\mathrm{cov}\left(rg_{\Theta}, rg_{\hat{\Theta}}\right)}{s_{rg_{\Theta}} \cdot s_{rg_{\hat{\Theta}}}}$$

where rg_{Θ} and $rg_{\hat{\Theta}}$ are the ranks of Θ and $\hat{\Theta}$, respectively, and $cov\left(rg_{\Theta}, rg_{\hat{\Theta}}\right)$ is the covariance of the corresponding ranks, $s_{rg_{\Theta}}$ and $s_{rg_{\hat{\Theta}}}$ are the sample standard deviations of rg_{Θ} and $rg_{\hat{\Theta}}$, respectively.

•Abundance Recovery Rate (*ARR*)

ARR is the percentage of the estimation over the ground truth, which is calculated by

$$ARR_i = \frac{\hat{\theta}_i}{\theta_i} \times 100\%, \quad (i = 1, 2, \cdots, I)$$

An accurate abundance estimation should have an *ARR* value close to 100%.

•Median Relative Difference (*MRD*)

MRD is the median of the relative difference of abundance estimates among all transcript isoforms within a sample, which is calculated by

$$MRD = median\left\{\frac{\left|\theta_i - \hat{\theta}_i\right|}{\theta_i}, \quad (i = 1, 2, \cdots, I)\right\}$$

A small *MRD* value indicates the good performance of abundance estimation.

•Normalized Root Mean Square Error (*NRMSE*)

NRMSE provides a measure of the extent to which the one-to-one relationship deviates from a linear pattern. It can be calculated by

$$NRMSE = \frac{\sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(\theta_i - \hat{\theta}_i\right)^2}}{s_\Theta}$$

where s_Θ is the sample standard deviation of Θ.

A good performance of abundance estimation should have a small value of *NRMSE*.

In the case of LRGASP, the above metrics can be calculated with simulated data and SIRVs.

Multiple replicates under two different conditions (ground truth is available)

Denote $\widehat{\theta}_{ijk}$ and θ_{ijk} as the estimation and ground truth of transcript isoform $i\,(i = 1, 2, \cdots, I)$ in a sample, where $j(j = 1, 2)$ represents different groups (i.e., conditions or tissues) and $k\,(k = 1, 2, \cdots, K)$ represents different replicates within the group j.

We assess the quantification performance by ROC (receiver operating characteristic) analysis of identifying true differentially expressed transcript isoforms. At first, we define Average Log Fold Change (*ALFC*) of transcript isoform i as:

$$ALFC_i = log\left(\frac{\frac{1}{K_2}\sum_{k_2=1}^{K_2}\left(\theta_{i2k_2}+1\right)}{\frac{1}{K_1}\sum_{k_1=1}^{K_1}\left(\theta_{i2k_1}+1\right)}\right).$$

Next, based on the ground truth values and a given threshold (e.g., 1 as below), we can define whether a transcript isoform is truly differentially expressed or not:

Positives (truly differentially expressed)

$$T = \{i|\,|ALFC_i| \geq 1\}$$

Negatives (not truly differentially expressed)

$$F = \{i|\,|ALFC_i| < 1\}$$

Based on the estimated values, we can also obtain the “predicted positives” and “predicted negatives” with the same threshold. Therefore, we can identify “true positives”, “true negatives”, “false positives” and “false negatives” to calculate the ROC-based statistics, including precision, recall, accuracy, F1-score, AUC and pAUC, and also plot ROC **(Supplementary Fig. S2)**.

The above metrics will be used for SIRVs and a subset of isoforms whose abundances were experimentally determined. In the case of SIRV sequencing, we would not expect fold change differences in different conditions, as the SIRVs were spiked in at relatively the same concentration in all samples.

Multiple replicates under different conditions (without the ground truth)

For multiple replicates under different conditions without the ground truth, we can still evaluate a quantification method by the "goodness" of its statistical properties, including **reproducibility**, **consistency** and **resolution entropy** that is also calculated for single sample data (**Supplementary Figs. S3-S5**)

•Reproducibility

The reproducibility statistic characterizes the average standard deviation of abundance estimates among different replicates (**Supplementary Fig. S3**), which is calculated by

$$RM = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} s_{ij}^2}$$

Here, s_{ij} is the sample standard deviation of $log\left(\widehat{\theta}_{ijk}+1\right)(k=1,2,\ldots,K)$, which is calculated by

$$s_{ij} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(log\left(\widehat{\theta}_{ijk}+1\right)-u_{ij}\right)^2},$$

where

$$u_{ij} = \frac{1}{K}\sum_{k=1}^{K} log\left(\widehat{\theta}_{ijk}+1\right).$$

With a small value of this metric, the method has high reproducibility. We can also plot s_{ij} versus average abundance u_{ij} to examine how standard deviation changes with respect to the abundance and the area under the curve is calculated as a secondary statistic.

•Consistency

A good quantification method tends to have the consistency of characterizing abundance patterns in different replicates. Here, we propose a consistency measure $C(\alpha)$ to examine the similarity of abundance profiles between mutual pairs of replicates (**Supplementary Fig. S4**), which is defined as:

$$C(\alpha) = \frac{1}{IJ \cdot C_K^2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{1 \le k_1 < k_2 \le K}^{I} P\left(\left\{ log\left(\widehat{\theta}_{ijk_1} + 1\right) < \alpha, log\left(\widehat{\theta}_{ijk_2} + 1\right) < \alpha \right\} \text{ or } \left\{ log\left(\widehat{\theta}_{ijk_1} + 1\right) \ge \alpha, log\left(\widehat{\theta}_{ijk_2} + 1\right) \ge \alpha \right\} \right),$$

where α is a customized threshold defining whether a transcript is expressed or not.

•Resolution Entropy (*RE*)

A good quantification method should have a high resolution of abundance values. For a given sample, a Resolution Entropy (*RE*) statistic characterizes the resolution of abundance estimation (**Supplementary Fig. S5**):

$$RE = -\sum_{m=1}^{M} P_m In\left(P_m\right), \text{where } P_m = \frac{n_m}{\sum_{j=1}^{M} n_j}.$$

Here, the abundance estimates are binned into M groups, where n_m represents the number of transcript isoforms with the abundance estimate $\widehat{\Theta} \in [m \cdot \alpha, (m+1) \cdot \alpha)$, and $\alpha = max\left(\widehat{\Theta}\right)/M$. $RE = 0$ if all transcript isoforms have the same estimated abundance values, while it obtains a large value when the estimates are uniformly distributed among M groups.

Evaluation with respect to multiple transcript features

Quantification performance could be influenced by different transcript features, such as exon-isoform structure and the true abundance level. Thus, we also evaluate the quantification performance for different sets of genes/transcripts grouped by transcript features, including number of isoforms, number of exons, ground truth abundance values and a customized statistic K-value representing the complexity of exon-isoform structures.

• K-value

Most methods for transcript isoform quantification assign sequencing coverage to isoforms; therefore, the exon-isoform structure of a gene is a key factor influencing quantification accuracy. Here, we use a statistic K-value (manuscript in preparation, **Supplementary Fig. S6**) to measure the complexity of exon-isoform structures for each gene. Suppose a gene of interest

has I transcript isoforms and E exons, and define $A = (a_{ie}), (i = 1, 2, \cdots, I; e = 1, 2, \cdots, E)$ as the exon-isoform binary matrix, where

$$a_{ie} = \begin{cases} 1, \text{if the isoform } i \text{ includes the exon } e \\ 0, \text{otherwise} \end{cases}$$

K-value is the condition number of the exon-isoform binary matrix A, which is calculated by

$$\text{K-value} = \frac{\sigma_{max}(A)}{\sigma_{min}(A)},$$

where $\sigma_{max}(A)$ and $\sigma_{min}(A)$ are the maximum and minimum singular values of the matrix A, respectively.

With genes binned by the complexity of their transcripts, we are also able to evaluate how often the rank of isoforms from highest to lowest abundance agree between different tools, regardless of a ground truth. In particular, we can evaluate how often the most abundant isoform (major isoform) has the same transcript structure as other methods and how this compares to the ground truth, if known. We would expect more variability in what is considered the major isoform of a gene that is correlated with an increased K-value.

Challenge 3 Evaluation: De novo transcript isoform detection without a high-quality genome

Challenge 3 will evaluate the applicability of lrRNA-seq for *de novo* delineation of transcriptomes in non-model organisms. The evaluation will assess the capacity of technologies and analysis pipelines for both defining accurate transcript models and for correctly identifying the complexity of expressed transcripts at genomic loci, when genome information is limited. We will evaluate two different scenarios: a) availability of a genome sequence but no gene annotation is available, and b) no genome assembly is available at all.

The challenge includes three types of datasets. The mouse ES transcriptome data (**Table 1**) will be used to request the reconstruction of mouse transcripts without making use of the available genome or transcriptome resources for this species. Models will be compared to the true set of annotations with the same set of parameters as in Challenge 1. While this dataset allows for a quantitative evaluation of transcript predictions in Challenge 3, it might deliver unrealistic results if analysis pipelines were somehow biased by information derived from prior knowledge of the mouse genome. To avoid this problem, a second dataset is used that corresponds to the whole

blood transcriptome of the Floridian manatee (*Trichechus matatus*). An Illumina draft genome of this organism exists (https://www.ncbi.nlm.nih.gov/assembly/GCF_000243295.1/) and the LRGASP consortium has generated a long-read genome assembly to support transcript predictions for this species. Additionally, Illumina data has been generated for this challenge and an existing set of 454 transcriptome data will be used. Again, we will evaluate pipelines that obtain transcript models without genome annotation but with these draft genome sequences, and without genome assembly data at all. Since no curated gene models exist for the manatee, Challenge 1 metrics cannot be applied. Instead, the evaluation of this dataset will involve comparative assessment of the reconstructed transcriptomes and experimental validation. For comparative assessment the following parameters will be calculated.

a. Total number of transcripts
b. Mapping rate of transcripts to the draft genomes (for pipelines not using genome data)
c. Length of the transcript models
d. % of transcripts with predicted coding potential
e. Does the pipeline provide gene/loci predictions? If yes, number of transcripts/loci
f. BUSCO completeness
g. % transcripts with Blast2GO annotation.
h. % of junctions with Illumina coverage
i. % junctions and transcripts with non-canonical splicing

We expect that good-performing pipelines will obtain longer transcripts, well supported by Illumina data, with high mapping rate to the draft genomes, most of them coding, and with higher BUSCO completeness and Blast2GO annotation potential.

Finally, the manatee long reads data also contain spiked-in SIRVs, which will be used to compute performance metrics for Challenge 3 analysis settings, using the same type of metrics as described for Challenge 1.

We will compare metric statistics across analysis pipelines and sequencing platforms. A number of genes will be selected for PCR-based experimental validation (see below), including transcripts of cytokine genes, which have been studied by LRGASP consortium members in detail[32]. Ferrante et al.[32] designed and validated primers to measure cytokine transcript levels in Florida manatees from blood samples, specifically for interleukin (IL)-2, -6, -10, interferon-

gamma (INF-gamma) and Tumor necrosis-alpha (TNF-alpha), and these methods will be adopted for development of isoform-specific assays.

Experimental validation of transcript models and expression estimates

Independent experimental validation will be performed to assess the accuracy of novel features and transcript isoforms characterized from the lrRNA-seq data from all challenges. In the evaluation of full-length transcripts, several local and long-range elements must be considered. Local elements include the 5' end of the transcript, splice site, junctions, novel exons, retained introns, and polyA sites. Long-range elements include chained series of junctions. We will employ a suite of several assays in order to validate both the local and long-range elements.

Challenge 1 Evaluation: Transcript isoform detection

The goal of this challenge is to assess the comprehensive and reliable detection of all transcripts in biological samples. Similar to past studies that have employed lrRNA-seq approaches towards characterizing the transcriptome, we expect that participants for this challenge will produce a large number of novel isoforms. Therefore, the approaches to assess the accuracy of transcript isoforms that were previously described (e.g., SIRV standards, GENCODE manual annotation) will be complemented with experimental validation.

We will employ several high-throughput sequencing-based assays to validate local elements, such as novel 5' ends, splice junctions, and polyA sites, on a "global" scale. Note that these experimental assays have or will be carried out using the same aliquot of total RNA as was used to generate the LRGASP datasets, minimizing differences in detected features due to biological or inter-laboratory variability. To validate novel 5' ends, we will use a recently generated a deep coverage CAGE data on the WTC-11 line. To validate novel splice junctions, we will also use Illumina RNA-seq to validate novel junctions and, wherever possible, exons or series of connected exons. To validate novel polyadenylation sites, we will collect polyA-seq data using the Quant-Seq method from Lexogen, which can map polyA sites *de novo*. Additionally, in select cases, novel 5' ends will be further corroborated through chromatin-based functional information derived from ENCODE data, such as the presence of PolII or histone marks that are indicative of active promoters.

Longer-range features within a transcript, such as chains of junctions, are difficult and sometimes impossible to detect through short-read sequencing approaches or traditional qPCR;

therefore, we will employ targeted amplicon sequencing followed by ONT, PacBio, and Sanger sequencing.

We plan to select 96 targets from human WTC-11 cells and 96 targets from the mouse 129/Casteneus cells. Each target will comprise a sequence region 300 to 1500 bp long. Two replicates each from the WTC-11 and 129/Casteneus sample will be apportioned for a reverse-transcriptase reaction followed by target amplification using isoform-specific primers. We will conduct the assay in plate format to allow for high-throughput processing. All products following RT-PCR will be pooled and subjected to long-read sequencing for validation. A subset of these samples will be selected for Sanger sequencing. Table 6 shows the breakdown of targets we will select.

Category	**WTC-11 (Human)**	**129/Casteneus (Mouse)**
Positive control	12	12
Negative control	12	12
Novel – detected in all platforms	12	12
ONT-specific	12	12
PacBio-specific	12	12
Miscellaneous category (e.g., bioinformatic pipeline-specific, intron retention, template switch artifact prediction, non-canonical splicing)	24	24

Table 6: Plan for targeted amplicon sequencing to validate novel junction chains in the LRGASP submissions.

Positive controls will be selected as subsegments of isoforms which are found in GENCODE human v39 and mouse vM28, all long-read datasets across the ONT and PacBio platforms, and a majority (>50%) of the computational pipelines. Negative controls will also be selected, which

would involve isoforms that are detected in other human and mouse cell types (e.g., pancreas cells), but for which there is no evidence of expression across any of the long-read datasets in LRGASP.

An open question in the field is the accuracy of novel isoforms that are frequently detected on long-read platforms, and so we will devote substantial effort towards validation of novel isoforms. At least 12 targets will involve junction chains that are novel (not in GENCODE) but found across all lrRNA-seq library types. We also reserve resources to validate platform-specific isoforms, in case they should arise. And, lastly, we reserve at least 24 targets for miscellaneous categories, such as if there is the appearance of certain isoforms in specific computational pipelines.

For novel target selection, preference will be given to select targets that correspond to the pre-selected 50 loci that will be manually annotated by GENCODE, and there will be close coordination between the working groups.

In addition to the validation using a PCR-based approach (Table 6), high-throughput validation of full-length transcripts will be obtained by application of the NRCeq strategy[25] on WTC-11 cells, which does not rely on PCR. NRCeq employs a chemical labeling strategy to add a signature oligonucleotide exclusively to the 5' caps of mRNAs, thus, full-length mRNA sequences from the 5' cap to the polyA sequence may be distinguished from incomplete sequence fragments. We will compare NRCeq data generated WTC-11 against models submitted by participants.

Challenge 2 Evaluation: Transcript isoform quantification

Challenge 2 involves the prediction of fold change in abundance at the gene and transcript isoform-level. For this purpose, the H1:H1-DE cell line mix will be compared to WTC11 cell line. H1 and WTC-11, both being stem cell lines, are expected to have similar expression patterns, but the H1:H1-DE mix would have gene and isoform expression more related to the definitive endoderm phenotype. To experimentally validate abundance changes, we will employ qPCR among isoforms of a gene which under altered expression as well as sequencing data on sample components before mixing.

qPCR of 10-20 transcript models will be performed. Due to the difficulty of properly resolving and apportioning signals for short junctions or exons to the full-length transcript isoforms they arose from, we will choose isoforms with low and high K-values, representing various levels of

identifiability. In some cases, we will increase the length of qPCR targets up to the 500-600 bp ranges so as to increase the resolution and specificity of isoform measurements. Internal standards will be spiked in for highest accuracy and precision of isoform abundance estimates.Targeted amplicon sequencing with long-read platforms will also be performed on these transcript models to determine fold-change differences.

Due to the challenges of isoform-level quantification and the lack of a gold standard, we devised a mixture sample, in which an undisclosed ratio of two samples is mixed before sequencing. For validation, we sequenced H1 and H1-DE samples individually to establish the isoforms present in only one or the other sample before mixing. In essence, the pre-mixed sample represents the "ground truth" of isoform expression before the mix. After the close of LRGASP submissions, the H1 and H1-DE long-read data will be released. Participants of Challenge 2, will need to provide transcript quantification from these additional datasets. Libraries and computational pipelines can then be evaluated based on how well the transcript quantification in the H1:H1-DE mix sample represents the expected ratios determined from quantification from the individual cell lines.

Challenge 3 Evaluation: De-novo transcript isoform detection without a high-quality genome

Similarly to Challenge 1, the primary goal of experimental validation in this challenge is to confirm the identity of *de novo* assembled isoforms, of which many will be novel.

A number of loci from well-studied immune-related genes will be selected for experimental PCR validation as in the mouse/human data.

To validate isoforms containing novel junction chains, we will employ a similar amplicon sequencing strategy as described in Challenge 1, in which up to 96 primer pairs will be used to amplify isoform-specific regions for subsequent detection on a sequencing platform.

In addition, there exists 454 sequencing data from these same samples which can also be leveraged for orthogonal validation.

Challenge submissions and timeline

Participants will submit challenge predictions on Synapse (https://www.synapse.org/#!Synapse:syn25007472).

The following is an overview of the data used for each challenge and the result files that will be submitted (**Supplementary Figure S7**).

- Challenge 1: transcript isoform detection with a high-quality genome (iso_detect_ref)
    - Samples
        - WTC11 (human iPSC cell line)
        - H1_mix (human H1 ES cell line mixed with human Definitive Endoderm derived from H1)
        - ES (mouse ES cell line)
        - human_simulation - simulated human reads (Illumina, ONT, and PacBio cDNA)
        - mouse_simulation - simulated mouse reads (Illumina and PacBio cDNA, ONT dRNA)
    - Result files:
        - models.gtf.gz
        - read_model_map.tsv.gz
- Challenge 2: transcript isoform quantification (iso_quant)
    - Samples
        - WTC11 (human iPSC cell line)
        - H1_mix (human H1 ES cell line mixed with human Definitive Endoderm derived from H1)
        - human_simulation - simulated human reads (Illumina, ONT, and PacBio cDNA)
        - mouse_simulation - simulated mouse reads (Illumina and PacBio cDNA, ONT dRNA)
    - Result files:
        - expression.tsv.gz
        - models.gtf.gz
- Challenge 3: de novo transcript isoform detection (iso_detect_de_novo)
    - Samples
        - Manatee (manatee whole blood)
        - ES (mouse ES cell line)
    - Result files:
        - rna.fasta.gz
        - read_model_map.tsv.gz

A submission to a challenge is an entry, consisting of one or more experiments. Each entry must meet the following requirements:

Requirements for Challenge 1 and 2

At least one experiment must be supplied for each sample available for a given challenge. Human and mouse samples will have biological replicates that should be used for the entry.

A major goal of LRGASP is to assess the capabilities of long-read sequencing for transcriptome analysis and also how much improvement there is over short-read methods. Additionally, long-read computational pipelines vary in their use of only long-read data or if they incorporate additional data for transcript analysis. To facilitate comparisons between long-read and short-read methods and variation in tool parameters, we break down submissions into different categories:

- long-only - Use only LGRASP-provided long-read RNA-Seq data from a single sample, library preparation method and sequencing platform.
- short-only - Use only LGRASP-provided short-read Illumina RNA-Seq data from a single sample. This is to compare with long-read approaches
- long and short - Use only LGRASP-provided long-read and short-read RNA-Seq data from a single long-read library preparation method and the Illumina platform. Additional accessioned data in public genomics data repositories can also be used.
- kitchen sink - Any combination of at least one LRGASP data set as well as any other accessioned data in public genomics data repositories. For example, multiple library methods can be combined (e.g. PacBio cDNA + PacBio CapTrap, ONT cDNA + ONT CapTrap+ ONT R2C2+ ONT dRNA, all data, etc.).

In all the above categories, the genome and transcriptome references specified by LRGASP should be used. For the long and short and kitchen sink category, additional transcriptome references can be used.

All replicates must be used in each experiment. Challenge 2 must report replicates separately in the expression matrix. Each team can only submit one entry per category.

For Challenge 1, the submitted GTF file should only contain transcripts that have been assigned a read. For Challenge 2, submitters have the option of quantifying against the reference

transcriptome or a transcriptome derived from the data (i.e., results from Challenge 1). The GTF used for quantification is included as part of the Challenge 2 submission.

The type of platform and libraries preparation method used in a given experiment, except for kitchen sink experiments, is limited to data from a single library preparation method plus sequencing technology (long-only). LRGASP Illumina short-read data of the same sample may optionally be used in an experiment with the LRGASP long-read data (long and short)

- Illumina cDNA - short-only
- Pacbio cDNA - long-only or long and short
- Pacbio CapTrap - long-only or long and short
- ONT cDNA - long-only or long and short
- ONT CapTrap - long-only or long and short
- ONT R2C2 - long-only or long and short
- ONT dRNA - long-only or long and short

Requirements for Challenge 3

At least one experiment must be supplied for each sample available for the challenge. Mouse samples will have biological replicates that should be used for the entry.

For similar reasons as described above, the data used for a given experiment must fit in one of the following categories:

- long-only - Use only LGRASP-provided long-read RNA-Seq data from a single sample, library preparation method and sequencing platform. No genome reference can be used.
- short-only - Use only LGRASP-provided short-read Illumina RNA-Seq data from a single sample. This is to compare with long-read approaches. No genome reference can be used.
- long and short - Use only LGRASP-provided long-read and short-read RNA-Seq data from a single long-read library preparation method and the Illumina platform. No genome reference can be used.
- long and genome - Use only LGRASP-provided long-read RNA-Seq data from a single long-read library preparation method. A genome reference sequence can be used.
- kitchen sink - Any combination of at least one LRGASP data set as well as any other accessioned data in public genomics data repositories. For example, multiple library methods can be combined (e.g. PacBio cDNA + PacBio CapTrap, ONT cDNA + ONT CapTrap+ ONT R2C2+ ONT dRNA, all data, etc.).

In all the above categories, except for kitchen sink a transcriptome reference cannot be used. The submitted FASTA file should only contain transcripts that have been assigned a read. Each team can only submit one entry per category.

LRGASP biological data is currently available at the ENCODE DCC (https://www.encodeproject.org/search/?type=Experiment&internal_tags=LRGASP). The simulated data is available from Synapse (https://www.synapse.org/#!Synapse:syn25683370). The competition launched on May 1, 2021 and challenge submissions are expected to close on October 1, 2021.

LRGASP Data QC

Initial quality control (QC) metrics were determined for the LRGASP data (**Figure 4**). Reads (ONT cDNA, dRNA, CapTrap) or consensus reads (PacBio cDNA and CapTrap and ONT R2C2) were aligned to the human or mouse genome as appropriate using minimap2 with the following parameters: -ax splice --secondary=no -G 400k. For each data type, the reads and their resulting alignments in sam format were parsed for the following parameters:

1) Number of aligned reads
2) Number of aligned reads with adapters on both ends

   For ONT dRNA this is not applicable as this workflow does not attach an adapter to the 5' end of molecules. For ONT cDNA and CapTrap this percentage was determined by pyChopper. For all other data types, all provided reads are assumed to have adapters on both ends as the pre-processing pipelines (lima and C3POa) discard reads otherwise.
3) median read length

   measured by the number of aligned bases (matches or mismatches)
4) median accuracy

   measured by matches/(matches+mismatches+indels)),
5) Percent of aligned reads where the orientation of the reads as determined by 5' and 3' adapter sequences agrees with the direction of the read alignment

   determined by minimap2 through splice site context (calculated only for the subset of reads with splice alignments with the ts:A: flag in their sam entry),
6) Percent of reads originating from spike-in molecules

determined by alignment to the SIRVomeERCC fasta entry in the genome sequence files

7) Pearson correlation between replicates

determined by quantifying gene expression for each replicate and calculating the pearson r value based on those expression values.

Table 7: Summary statistics for LRGASP data. For each sample, replicates were combined when reporting statistics.

Sample	ES					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	SequelII	SequelII
# of Flowcells/SMRT cells	3	3	6	3	3	9
# of raw reads	4,325,200	59,746,818	7,862,883[1]	56,684,765	9,689,619	23,487,808
# of supplied reads	**3,975,725**	**57,055,583**	**5,930,487**	**50,697,997**	**5,090,848**	**8,733,814**
# of aligned reads	3,836,020	44,873,564	5,914,779	49,741,194	5,028,403	8,199,908
# of aligned reads with adapters	N/A	40,190,805	5,914,779	32,206,495	5,028,403	8,199,908
Median Read length	830	519	1,755	591	903	2,090
Median Identity (Q score)	9.8	12.7	18.6	12.3	21.3	20.9
% Directionality	99.54	98.59	99.74	94.66	99.88	99.55
% of spike-in reads	0.71	1.02	2.03	2.41	1.77	1.85
Pearson r2 (gene level)	0.99	0.99	0.98	0.99	0.98	0.97

[1]R2C2 libraries for ES and WTC11 libraries were multiplexed and raw reads cannot be demultiplexed directly. Raw read numbers for these libraries are therefore calculated based on the ES/WTC11 ratio of demultiplexed supplied consensus reads and total number of subreads.

Sample	WTC11					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	SequelII	SequelII
# of Flowcells/SMRT cells	3	3	6	3	3	9
# of raw reads	3,229,571	53,463,774	6,994,789[1]	56,730,485	13,463,712	28,567,150
# of supplied reads	**2,988,430**	**51,194,535**	**5,275,737**	**50,902,303**	**6,399,632**	**7,424,923**
# of aligned reads	2,931,482	43,085,527	5,271,334	49,930,350	6,304,610	7,373,147
# of aligned reads with adapters	N/A	37,275,068	5,271,334	31,348,191	6,304,610	7,373,147
Median Read length	854	610	1,802	564	864	2,209
Median Identity (Q score)	9.8	12.9	19.3	12.9	22.5	23.8
% Directionality	99.76	99.11	99.92	96.28	99.92	99.67
% of spike-in reads	0.6	1.45	2.27	2.79	2.26	2.25
Pearson r2 (gene level)	0.92	0.96	0.94	0.99	0.96	0.90

[1]R2C2 libraries for ES and WTC11 libraries were multiplexed and raw reads cannot be demultiplexed directly. Raw read numbers for these libraries are therefore calculated based on the ES/WTC11 ratio of demultiplexed supplied consensus reads and total number of subreads.

Sample	H1_mix					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	SequelII	SequelII
# of Flowcells/SMRT cells	3	3	6	3	3	6
# raw reads	4,223,164	55,927,828	7,093,671	54,055,468	10,534,880	24,290,762
# of supplied reads	**3,969,603**	**52,927,595**	**5,231,255**	**49,883,469**	**5,511,853**	**5,511,357**
# of aligned reads	3,905,742	43,026,016	5,229,686	48,424,901	5,436,170	5,480,635
# of aligned reads with adapters	N/A	36,653,422	5,229,686	28,099,080	5,436,170	5,480,635
Median Read length	891	619	1,782	604	1,036	2,376
Median Identity (Q score)	10.0	12	18.7	12.4	24.3	23.7
% Directionality	99.8	99.19	99.74	**76.15**[1]	99.91	99.63
% of spike-in reads	0.77	1.5	1.69	1.59	1.33	1.97
Pearson r2 (gene-level)	0.99	0.997	0.98	0.96	0.98	0.98

[1]Replicate 3 of the H1_mix sample appears to be an outlier among the CapTrap ONT library type. Replicates 1 and 2 show % directionality ~95% similar to what is observed in the other samples for this library type.

Sample	Manatee	Manatee
Method	cDNA	cDNA
Tech	ONT	PacBio
Platform	MinION	Sequel I + Sequel II
# of Flowcells/SMRT cells	3	1+3
# of supplied reads	**40,948,571**	**6,883,684**
# of aligned reads	32,833,840	6,877,181
# of aligned reads with adapters	27,381,394	6,877,181
Median Read length	540	894
Median Accuracy (Q score)	12.5	25.2
% Directionality	97.2	99.76
% of spike-in reads	**14.05***	**33.78***
*spike-in percentage is higher than expected		

Data and code availability

All code and documentation associated with the LRGASP Consortium can be found through https://www.gencodegenes.org/pages/LRGASP/ and https://github.com/LRGASP.

Acknowledgments

We thank Lexogen, Oxford Nanopore Technologies (ONT), and Pacific Biosciences for helpful discussions. ONT provided partial support of flow cells and reagents. We thank Xingjie Ren and Yin Shen for providing WTC11 cells, Takayo Sasaki and Dave Gilbert for providing the F121-9 hybrid mouse ES cells, and Alyssa Cousineau, Krishna Mohan Parsi, and Rene Maehr for providing human H1 and H1-DE cells. We also thank Mark Akeson and Miten Jain for providing resources and technical advice for Nanopore sequencing. We thank Julie Visser for contributing artwork that gives an overview of the LRGASP Consortium. The project is supported by the following grants: Pew Charitable Trust (A.N.B.), NIGMS R35GM138122(A.N.B.), NHGRI U41HG007234 (J.L, M.D., R.G. and S.C-S) and UM1 HG009443 (A.M. and B.W.), an institutional fund of the Department of Biomedical Informatics, The Ohio State University (K.F.A.), NHGRI R01HG008759 (K.F.A.), SPBU 73023672 (A.P). J.E.L., J.M.M. and A.F. are supported by National Human Genome Research Institute of the National Institutes of Health [U41HG007234]; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health; Wellcome Trust [WT108749/Z/15/Z, WT200990/Z/16/Z]; European Molecular Biology Laboratory. We acknowledge Ellie Schiller Homosassa Springs Park for providing archive Lorelei blood samples.

Competing Interests

Design of the project was discussed with Oxford Nanopore Technologies (ONT), Pacific Biosciences, and Lexogen. ONT provided partial support of flow cells and reagents. S.C-S and A.N.B. have received reimbursement for travel, accommodation and conference fees to speak at events organised by ONT.

1. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4821–30 (2013).
2. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
3. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100 (2017).
4. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
5. Byrne, A., Cole, C., Volden, R. & Vollmers, C. Realizing the potential of full-length transcriptome sequencing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190097 (2019).
6. Oikonomopoulos, S. *et al.* Methodologies for Transcript Profiling Using Long-Read Technologies. *Front. Genet.* **11**, 606 (2020).
7. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
8. Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A. & Tilgner, H. U. Getting the Entire Message: Progress in Isoform Sequencing. *Front. Genet.* **10**, 709 (2019).
9. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).
10. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
11. Reese, M. G. *et al.* Genome annotation assessment in Drosophila melanogaster. *Genome Res.* **10**, 483–501 (2000).
12. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7 Suppl 1**, S2.1–31 (2006).
13. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the

sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9726–9731 (2018).

14. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996).
15. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
16. Foote, A. D. *et al.* Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
17. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
18. Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303**, 19–44 (1999).
19. Shibata, Y. *et al.* Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *Biotechniques* **30**, 1250–1254 (2001).
20. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018) doi:10.1101/gr.222976.117.
21. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
22. Jo, J. *et al.* Midbrain-like Organoids from Human Pluripotent Stem Cells Contain Functional Dopaminergic and Neuromelanin-Producing Neurons. *Cell Stem Cell* **19**, 248–257 (2016).
23. Hafezqorani, S. *et al.* Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *Gigascience* **9**, (2020).
24. Takahashi, H., Nishiyori-Sueki, H., Ramilowski, J. A., Itoh, M. & Carninci, P. Low Quantity single strand CAGE (LQ-ssCAGE) maps regulatory enhancers and promoters. doi:10.1101/2020.08.04.231969.
25. Mulroney, L. *et al.* Identification of high confidence human poly(A) RNA isoform scaffolds

using nanopore sequencing. doi:10.1101/2020.11.18.389049.

26. Schulz, L. *et al.* Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol.* **22**, 190 (2021).
27. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
28. Searle, S. M. J., Gilbert, J., Iyer, V. & Clamp, M. The otter annotation system. *Genome Res.* **14**, 963–970 (2004).
29. Sonnhammer, E. L. & Durbin, R. A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* **10**, 301–307 (1994).
30. Sonnhammer, E. L. & Durbin, R. An expert system for processing sequence homology data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 363–368 (1994).
31. Sonnhammer, E. L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–10 (1995).
32. Ferrante, J. A., Hunter, M. E. & Wellehan, J. F. X. DEVELOPMENT AND VALIDATION OF QUANTITATIVE PCR ASSAYS TO MEASURE CYTOKINE TRANSCRIPT LEVELS IN THE FLORIDA MANATEE (TRICHECHUS MANATUS LATIROSTRIS). *J. Wildl. Dis.* **54**, 283–294 (2018).

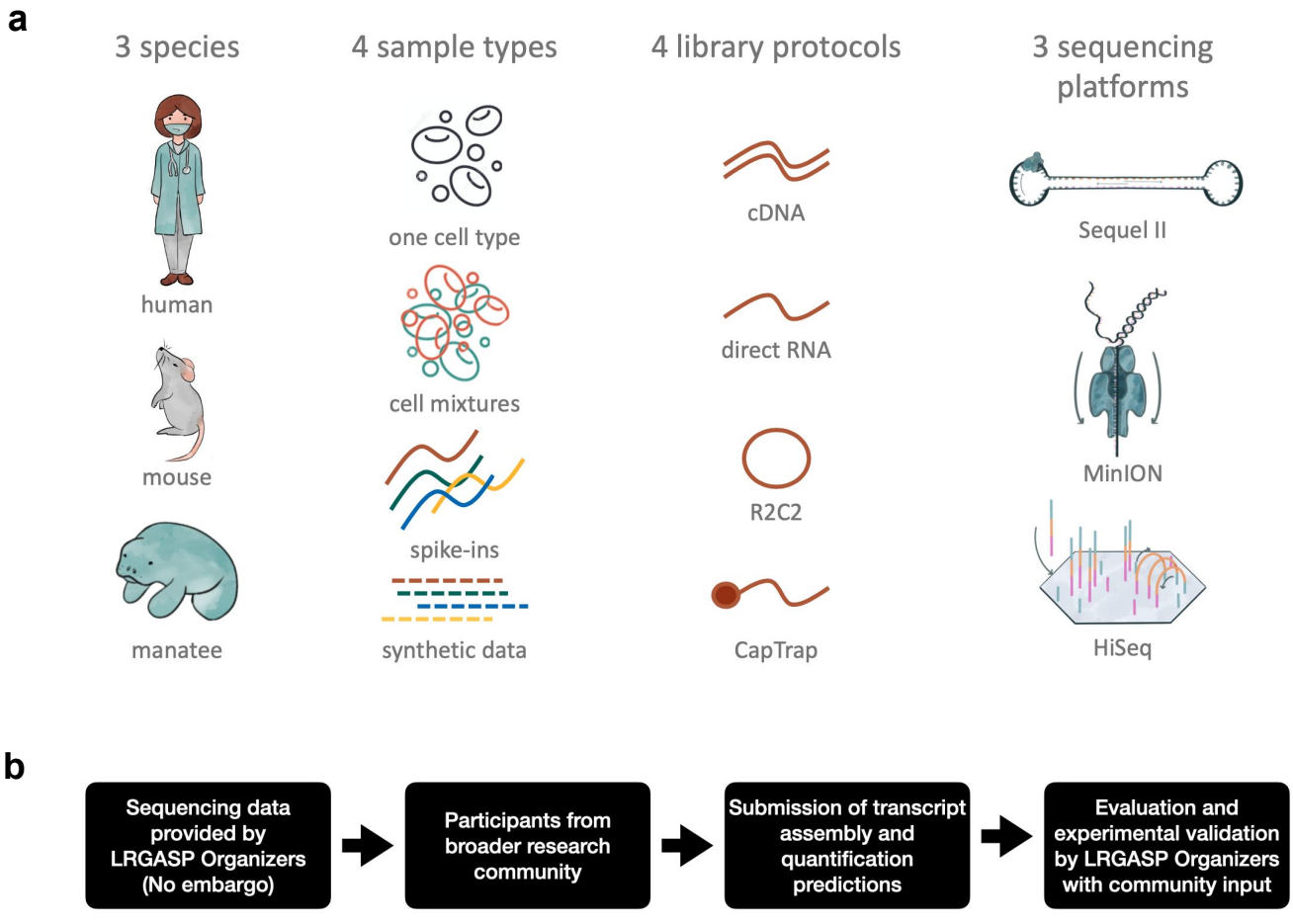

Fig. 1: Overview of the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP). a, LRGASP Consortium as a research community effort. **b**, Overview of LRGASP data.

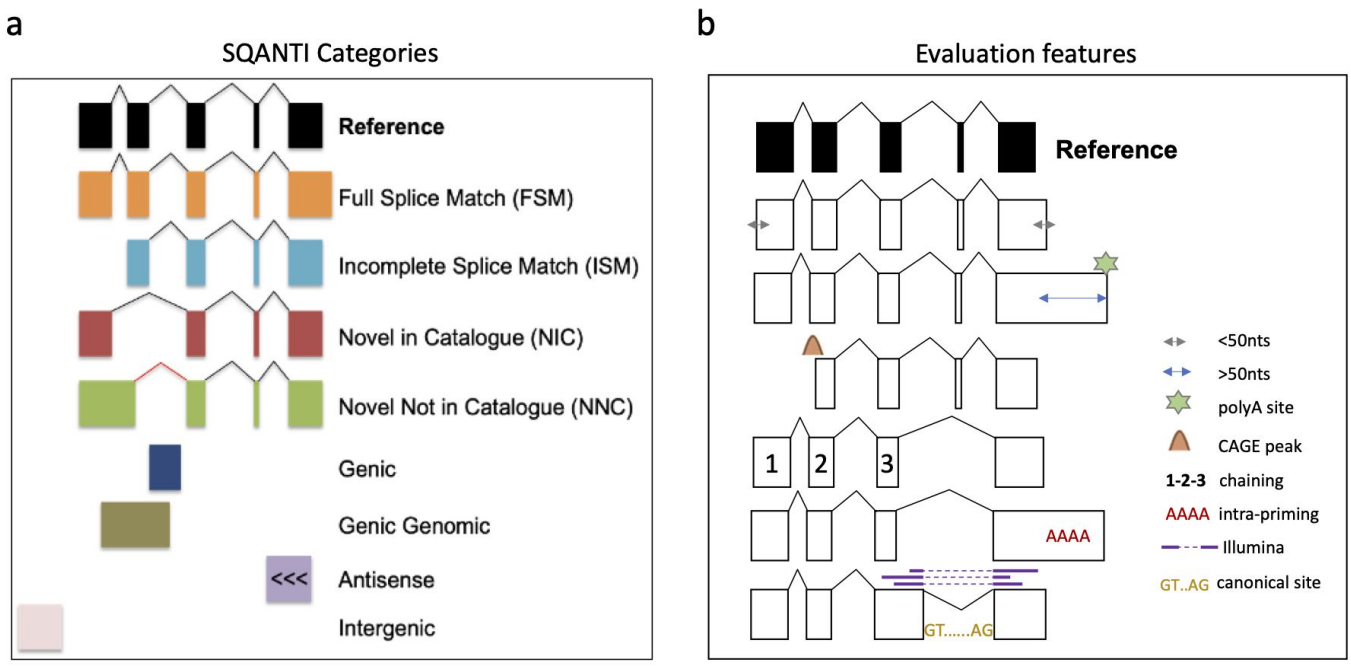

Fig. 2: SQANTI-based evaluation of transcript identification methods for Challenges 1 and 3. a, Transcripts are compared to a best matched reference transcript and categorized based on shared junctions between the reference. **b,** Additional features that are considered when evaluating transcript models

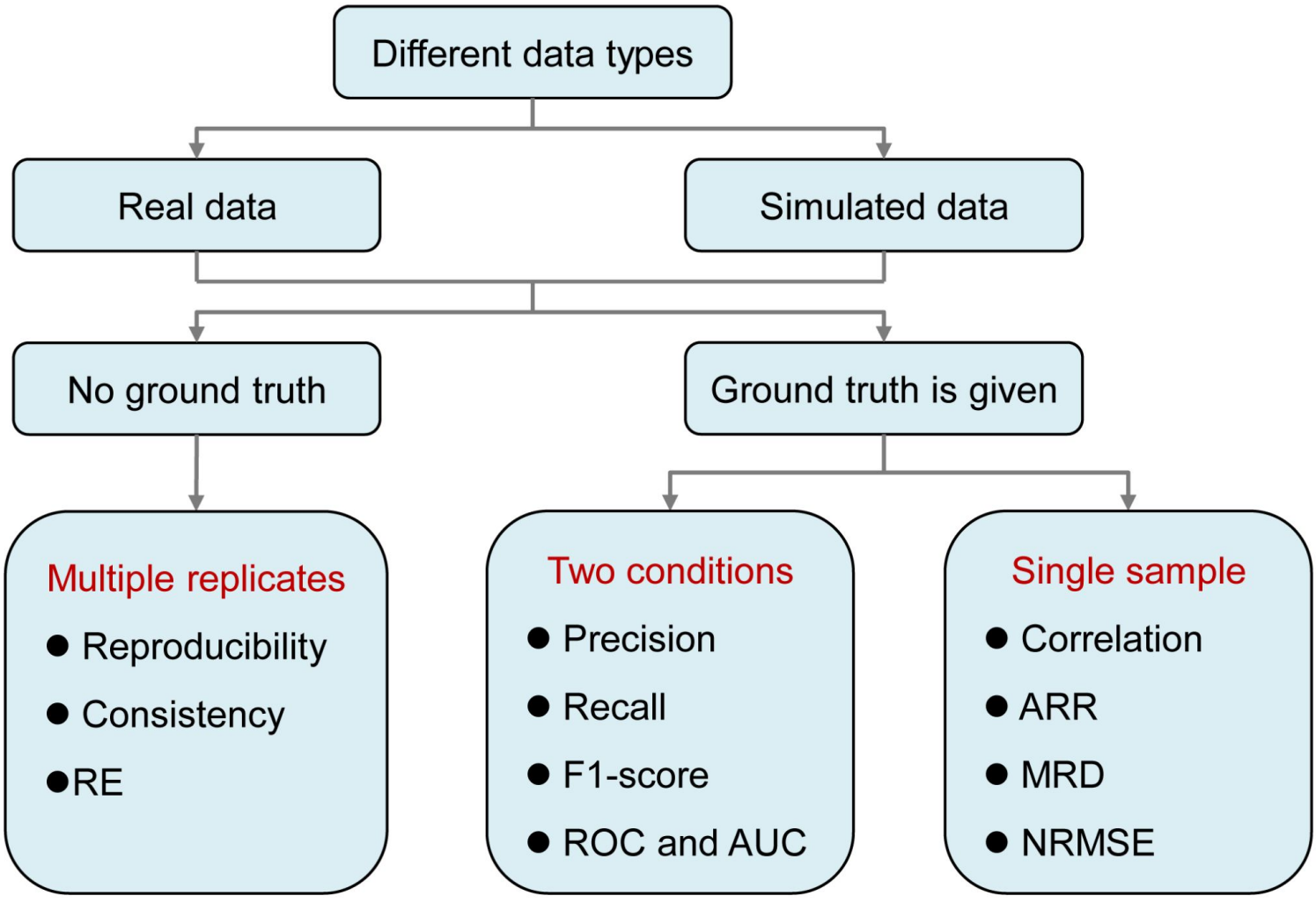

Fig. 3: Evaluation metrics of gene isoform quantification under different data types. RE - Resolution Entropy, ARR - Abundance Recovery Rate, MRD - Median Relative Difference, NRMSE - Normalized Root Mean Square Error

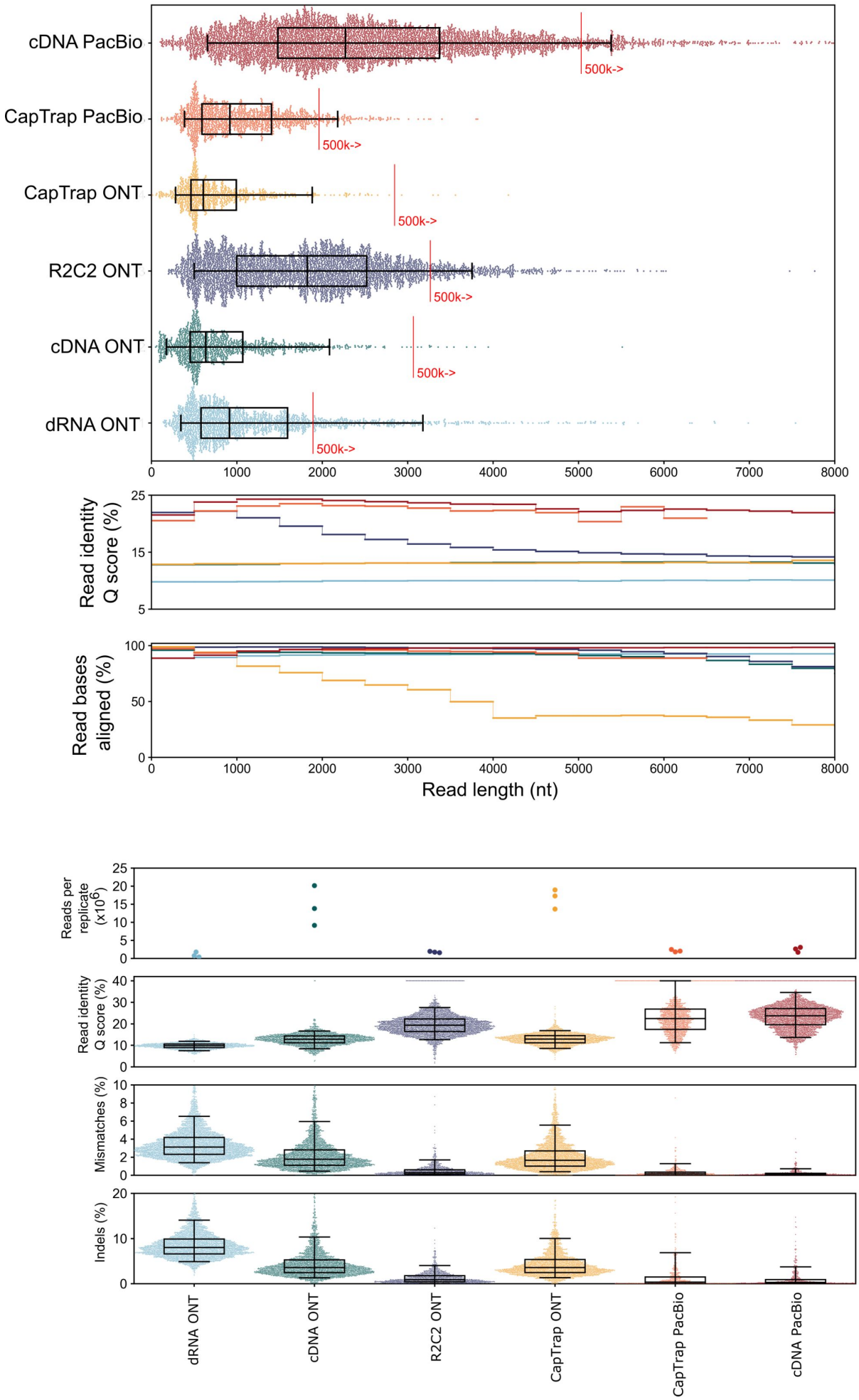

Fig. 4: Summary of LRGASP Data

Supplementary Files

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

- SupplementaryTable1.txt
- RegisteredReportSuppInfo210730wFigures.pdf
- RegisteredReportSuppInfo210730wFigures.pdf