

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

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 Marfa

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>

Angela N. Brooks (✉ anbrooks@ucsc.edu)

University of California, Santa Cruz <https://orcid.org/0000-0002-7898-3073>

Analysis

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

Posted Date: August 3rd, 2021

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

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Systematic assessment of long-read RNA-seq methods for transcript**

2 **identification and quantification**

3 Francisco J. Pardo-Palacios^{1,29}, Fairlie Reese^{2,3,29}, Sílvia Carbonell-Sala^{4,29}, Mark Diekhans^{5,29},
4 Cindy E. Liang^{6,29}, Dingjie Wang^{7,29}, Brian Williams^{8,29}, Matthew S. Adams⁶, Amit K. Behera⁹,
5 Julien Lagarde⁴, Haoran Li⁷, Andrey D. Pribelski¹⁰, Gabriela Balderrama-Gutierrez^{2,3},
6 Muhammed Hasan Çelik^{2,3}, Maite De María^{11,12}, Nancy Denslow¹³, Natàlia Garcia-Reyero¹⁴,
7 Stefan Goetz¹⁵, Margaret E. Hunter¹⁶, Jane E. Loveland¹⁷, Carlos Menor¹⁵, David Moraga¹⁸,
8 Jonathan M. Mudge¹⁷, Hazuki Takahashi¹⁹, Alison D. Tang⁹, Ingrid Ashley. Youngworth²⁰, Piero
9 Carninci^{19,21}, Roderic Guigó^{4,22}, Hagen U. Tilgner²³, Barbara J. Wold⁸, Christopher Vollmers^{9,30},
10 Gloria M. Sheynkman^{24,25,26,30}, Adam Frankish^{17,30}, Kin Fai Au^{7,30}, Ana Conesa^{27,28,30*}, Ali
11 Mortazavi^{2,3,30*}, Angela N. Brooks^{5,9,30*}

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

37 **Abstract**

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

56

57 **Introduction**

58 There is a growing trend of using long-read RNA-seq (lrRNA-seq) data for transcript
59 identification and quantification, primarily with Oxford Nanopore Technologies (ONT) and Pacific
60 Biosciences (PacBio) platforms¹⁻⁴. Consequently, there is a need to evaluate these approaches
61 for transcriptome analysis to compare the impact of different sequencing platforms, multiple
62 sequencing library preparation methods, and computational analysis methods (Reviewed in ⁵⁻⁸).

63

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

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

70

71 To evaluate long-read approaches for transcriptome analysis, we formed the Long-read RNA-
72 Seq Genome Annotation Assessment Project (LRGASP) Consortium modeled after the
73 previous GASP¹¹, EGASP¹², and RGASP^{9,10} efforts. For this project, we aim for an open
74 community effort in order to be as transparent and inclusive as possible in evaluating
75 technologies and computational methods (**Fig 1**).

76

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

84

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.

85 **Box 1: Overview of the LRGASP Challenges**

86

87 The LRGASP Challenges will use data produced by the LRGASP Consortium Organizers (**Fig**
88 **1b, Table 1, Supplementary Table 1**). The samples for Challenges 1 and 2 consist of human

89 and mouse ENCODE biosamples with extensive chromatin-level functional data generated
90 separately by the ENCODE Consortium. These include the human WTC-11 iPSC cell line and a
91 mouse 129/Casteneus ES cell line for Challenge 1 and a mix of H1 and Definitive Endoderm
92 derived from H1 (H1-DE) for Challenge 2. In addition, individual H1 and H1-DE samples are
93 being sequenced on all platforms; however, those reads will not be released until after the end
94 of the challenge. All samples were grown as biological triplicates with the RNA extracted at one
95 site, spiked with 5'-capped Spike-In RNA Variants (Lexogen SIRV-Set 4), and distributed to all
96 production groups. After sequencing, reads for human and mouse samples were deposited at
97 the ENCODE Data Coordination Center (DCC) for community access, including but not limited
98 to the challenges. A single replicate of manatee whole blood transcriptome was generated for
99 Challenge 3. For each sample, we performed different cDNA preparation methods, including an
100 early-access ONT cDNA kit (PCS110), ENCODE PacBio cDNA, R2C2¹³ for increased sequence
101 accuracy of ONT data, and CapTrap to enrich for 5'-capped RNAs. CapTrap is derived from the
102 CAGE technique¹⁴ and was adapted for lrrNA-seq (manuscript in preparation). We also
103 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
<i>Human H1 ES cell line</i>	3	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Human H1 Definitive Endoderm cell line</i>	3	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Trichechus manatus</i> peripheral blood mononuclear cells	1	Yes	Yes	No	No	No	No	Yes

104 **Table 1: Overview of LRGASP sequencing data.** The H1 and H1 Definitive Endoderm
105 samples are sequenced but are not available to participants until the close of challenges.
106
107 Participants may provide multiple submissions for each challenge (detailed in **Challenge**
108 **submissions and timeline**) and in any or all challenges. We will compare solutions where only
109 IrRNA-seq data was used and solutions that include additional publicly-available data.
110 Depending on the challenge, they will submit either a GTF or quantification file, additional
111 metadata, and a link to a repository (e.g., Github) where a working copy of the exact analysis
112 pipeline used to generate their results can be downloaded. We expect to re-run analysis
113 pipelines for well-performing submissions to help ensure reproducibility. The evaluation of the
114 challenge will comprise both bioinformatics and experimental approaches. SQANTI3
115 (<https://github.com/ConesaLab/SQANTI3>) will be used to obtain transcript features and
116 performance metrics that will be computed on the basis of SIRV-Set 4 spike-ins, simulated data,
117 and a set of undisclosed, manually curated transcript models defined by GENCODE¹⁵. Human
118 models will further be compared to histone modification CHIP-seq, open chromatin, CAGE, and
119 poly(A)-seq results. Experimental validation will be performed on a select number of loci with
120 either high agreement or disagreement between sequencing platforms or analysis pipelines.
121 Evaluation scripts and experimental protocols will be publicly available in advance of submission
122 deadlines (**Data and code availability**).

123 **Methods**

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

126 **Capping SIRVs**

127 Exogenous synthetic RNA references (spike-ins) are widely used to calibrate measurements in
128 RNA assays, but they lack the 7-Methylguanosine (m⁷G) cap structure that most natural
129 eukaryotic RNA transcripts bear at their 5' end. This characteristic makes commercial spike-in
130 mixes unsuitable for library preparation protocols involving 5' cap enrichment steps. Therefore,
131 we enzymatically added the appropriate m⁷G structure to the SIRV standards used in this
132 challenge. Specifically, the pp5'N structure present at the 5' end of spike-in sequence was used
133 as a template for the Vaccinia capping enzyme (catalog num M2080S, New England BioLabs)
134 to add the m⁷G structure to SIRV-Set 4 (Iso Mix E0 / ERCC / Long SIRVs, catalog num 141.03,

135 Lexogen). A total of ten vials of SIRV-Set 4 (100 µl) were employed to perform the capping
136 reaction (final total mass of 535 ng). The reaction was performed following the
137 recommendations of the manufacturer's capping protocol with two minor changes: 3.5 µl of
138 RNase inhibitors (RNasin Plus RNase Inhibitor, catalog num N2611, Promega) were added to
139 the capping reaction to avoid RNase degradation, and the incubation time was extended from
140 30 minutes to two hours, following a recommendation from New England BioLabs technical
141 support scientists. The final capping reaction was purified by using 1.8x AMPure RNA Clean XP
142 beads (catalog num. A63987, Beckman Coulter) and resuspended in 100 µl of nuclease-free
143 water.

144 **Mouse and human RNA sample preparation**

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

155

156 **Manatee RNA sample preparation**

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

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

181

182 **Manatee genome sample preparation**

183 The genome of the Florida manatee Lorelei was sequenced using Nanopore and Pacbio.
184 Lorelei is the same individual manatee for which an Illumina-based genome assembly was
185 released by the Broad Institute in 2012¹⁶. An EDTA, -80°C whole blood sample aliquot was
186 used. gDNA was extracted from 1400 µl of blood using the DNeasy kit (QIAGEN, MD, USA)
187 following the companies' specifications for 100 µl aliquots of blood. Thawed blood was diluted
188 1:1 with RNA free Phosphate buffered saline 1x (Gibco, UK), 20 µl of proteinase K (QIAGEN,
189 MD, USA), and 200 ul of AL lysis buffer (QIAGEN, MD, USA) and vortexed immediately. It was
190 incubated at 56 °C for 10 minutes. Then, we added 200 µl of ethanol 96% and mixed it
191 thoroughly. The mixture was added to the DNeasy mini spin-column and centrifuged at 6,000 x
192 g for 1 minute. The column was washed with 500 µl of AW1 solution (QIAGEN, MD, USA) and
193 centrifuged at 6,000 x g for 1 minute and followed with a wash with 500 µl AW2 (QIAGEN, MD,
194 USA) and centrifuged 20,000 x g for 3 minutes. gDNA was eluted twice with 100 µl of AE
195 buffer added to the center of the column, incubated for 1 minute, and centrifuged 6,000 x g for 1
196 minute. The first and second elution from the DNeasy mini spin-column were pooled and
197 concentrated using a speed vacuum for 20 minutes in which each preparation was reduced
198 from 200 to 50 µl. All gDNA tubes were pooled and the DNA was cleaned with AM Pure
199 magnetic beads (Beckman Coulter-Life Sciences, IN, USA) at a ratio of 0.5:1, beads volume to

200 gDNA volume (50 µl of beads to 100 µl of gDNA). gDNA bound to the beads was washed twice
201 with 1 ml of 70% ethanol. Ethanol traces were removed by quick spin to the bottom of the tube
202 and removed with a pipette. Then, the beads were dried for 2 minutes and gDNA was eluded in
203 55 µl of EB buffer (QIAGEN, MD, USA) at 37 °C with 10 minutes of incubation. This process
204 was repeated twice. Quantification of gDNA was performed with a Qubit™ fluorometer
205 (Thermo Fisher Scientific) and the quality of the gDNA was assessed using a Genomic Agilent
206 TapeStation (Santa Clara, CA, USA). The final DNA quantity was 28.8 µg of DNA at a
207 concentration of 267 ng/µl. The DNA Integrity Number (DIN) was 8.8 and the peak size was
208 54.5 kb.

209

210 **cDNA preparation for Illumina and PacBio sequencing of human and mouse**

211 PacBio cDNA synthesis was performed using a modified version of the Picelli protocol¹⁷ with the
212 Maxima H- reverse transcriptase. Total RNA was treated with exonuclease to remove
213 transcripts without a cap. 2 µl of exonuclease-treated RNA were mixed with a priming reaction
214 (RNAse inhibitor, dNTP's and water) was incubated at 72°C for 3 minutes, then ramps down to
215 50°C. While in the PCR block we added oligo dT (stock concentration 10 nM) and were
216 incubated 3 min at 50°C. We then added a first strand synthesis buffer (5x RT buffer, TSOligo,
217 water) that had previously been incubated at 50°C for one minute. The previous reaction was
218 then incubated in the PCR block (Extension at 50°C for 90 min, 85°C for 5 min and held at 4°C).
219 To the same reaction we added a mix for amplification (2x reaction buffer, IS primers - 20 nM
220 stock, water and SeqAmp polymerase). Then we ran a PCR program to amplify the cDNA (95°C
221 1 min, 98°C 15 sec, 65°C 30 sec and 68°C 13 min. The cycle repeats 10 times, which is
222 followed by incubation at 72°C for 10 min and holding at 4°C. The amplified products were
223 purified using SPRI beads and checked for quality in a bioanalyzer.

224

225 **PacBio library preparation of human and mouse libraries**

226 To build PacBio libraries, we started from 500 ng of polyA selected cDNA. The ends of the
227 cDNA were repaired first in order for the cDNA molecule to be suitable for ligation of SMRTbell
228 adapters. We added a damage repair reaction (DNA prep buffer, NAD and DNA damage repair)
229 and then incubated at 37°C for 30 min. Then End prep mix was added and incubated at 20°C for
230 30 min and 65°C 20 min. Ligation of the adapter at the ends of the cDNA was done by adding a
231 ligation mix (pacbio adapters, ligation mix, ligation enhancer and ligation additive), then it was

232 incubated at 20°C for 60 min. Final libraries were cleaned up using SPRI beads and we
233 recorded the size and concentration of samples. Once the ligation step was done and the
234 libraries passed the QC, a sequencing primer was annealed to the adapters in the UCI GHTF
235 sequencing facility to allow for the binding of the polymerase during sequencing.
236

237 **CapTrap preparation for PacBio and ONT sequencing of human and mouse**

238 CapTrap is a technique developed by the Guigó laboratory (CRG, Barcelona, Spain) in
239 collaboration with the group of Piero Carninci in RIKEN, Japan. The method enriches for full-
240 length transcripts by selection of the 7-Methylguanosine (m7G) cap structure present at the 5'
241 ends of RNA transcripts, followed by specific cap- and polyA- dependent linker ligations. The
242 cDNA libraries generated using this method are compatible with long-read sequencing platforms
243 (ONT or PacBio). The protocol starts with first strand synthesis (PrimeScript II Reverse
244 Transcriptase, catalog num. 2690A, Takara) where 5 µg of total RNA polyA+ RNAs are fully
245 reverse transcribed using a 16-mer anchored dT oligonucleotide. First strand synthesis was
246 performed at 42 °C for 60 minutes. Resulting products were purified with 1.8x AMPure RNA Clean
247 XP beads (catalog num. A63987, Beckman Coulter). After the first-strand generation, the m7G
248 cap structure at the 5' end of the transcripts is selectively captured using the CAP-trapper
249 technique^{14,18}, which leads to the removal of uncapped RNAs. The diol group on the m⁷G cap is
250 oxidized with 1M NaOAc (pH 4.5) and NaIO₄ (250 mM). Tris HCl (1M, pH 8.5) was added to stop
251 the reaction and the whole reaction was purified with 1.8x AMPure RNA Clean XP beads.
252 Aldehyde groups were biotinylated using a mixture containing NaOAc (1M, pH 6.0) and Biotin
253 (Long Arm) Hydrazide (100 mM, catalog num. SP-1100, Vector Laboratories). The resulting
254 mixture was then incubated for 30 minutes at 40°C and purified with 1.8x AMPure RNA Clean XP
255 beads. Single strand RNA was degraded by RNase ONE Ribonuclease (catalog num. M4261,
256 Promega) for 30 minutes at 37°C and purified with 1.8x AMPure RNA Clean XP beads. The m7G
257 cap structure bound to biotin is then selected using M-270 streptavidin magnetic beads (catalog
258 num. 65305, Thermo Fisher Scientific). M-270 streptavidin magnetic beads were equilibrated with
259 CapTrap Lithium chloride/Tween 20 based binding buffer. Sample recovered after RNase ONE
260 purification was bound to equilibrated M-270 streptavidin magnetic beads (incubation at 37°C for
261 15 minutes), washed 3 times with CapTrap Tween20 based washing buffer and released by heat
262 shock for 5 minutes at 95°C and quickly cooled on ice. A second release was performed, and the
263 supernatant was also collected and mixed with the eluate from the previous release. The released
264 sample was treated with RNase H (60 U/µl, Ribonuclease H <RNase H>, catalog num. 2150,

265 Takara), RNase ONE (10 U/μl) and CapTrap release buffer (incubated at 37°C for 30 minutes),
266 purified with 1.8x AMPure XP beads (catalog num. A63881, Beckman Coulter) and concentrated
267 by using a speed vac. After this cap specific selection, two double-stranded linkers, carrying a
268 unique molecular identifier (UMI), are specifically ligated to the first strand cDNA¹⁹. Linker ligation
269 (DNA Ligation Kit <Mighty Mix>, catalog num. 6023, Takara) was performed in two separate
270 steps. First the 5' linker was ligated, purified twice, to completely eliminate the non-incorporated
271 linkers, with 1.8x AMPure XP beads and concentrated by using a speed vac. Then the 3' linker
272 was ligated, purified once with 1.8x AMPure XP beads and finally concentrated by using a speed
273 vac. The double stranded linkers are converted into single strand by Shrimp Alkaline Phosphatase
274 (1 U/μl SAP, catalog num. 78390, Affymetrix) and Uracil-Specific Excision Reagent (1 U/μl USER,
275 catalog num. M5505L, NEB) treatment. This reaction was incubated for 30 minutes at 37°C, 5
276 minutes at 95°C and finally placed on ice. The sample was then purified with 1.8x AMPure XP
277 beads. After this treatment, the two linkers which serve as priming sites for the polymerase (2x
278 HiFi KAPA mix, catalog num. 7958927001-KK2601, Kapa), enable the synthesis of the full-length
279 second strand. The mixture was incubated for 5 minutes at 95°C, 5 minutes at 55°C, 30 minutes
280 at 72°C and finally held at 4°C until 1 μl Exonuclease I (20U/μl, catalog num. M0293S, NEB) was
281 added to each sample. The sample was then incubated for 30 minutes at 37°C and afterwards,
282 purified twice with 1.8x and 1.4x (respectively) AMPure XP beads and finally concentrated in a
283 speed vac. The resulting cDNA is amplified (TaKaRa LA Taq, catalog num. RR002M, Takara) via
284 long and accurate PCR (LA PCR) protocol. In order to avoid PCR duplicates, each sample was
285 split in two PCR independent reactions and amplified 16 cycles with 15 seconds at 55°C for
286 annealing, and 8 minutes at 65°C for extension. The 2 PCR replicates were merged and purified
287 with 1x AMPure XP beads. Samples were quantified with Qubit (Qubit 4 Fluorometer, Thermo
288 Fisher Scientific) and quality-checked with BioAnalyzer (Agilent 2100 Bioanalyzer, Agilent
289 Technologies).

290
291 CapTrap MinION cDNA sequencing was performed with 500 ng of cDNA sample coming from
292 CapTrap cDNA protocol and strictly following the SQK-LSK109 adapter ligation protocol (ONT).
293 The cDNA sequencing on MinION platform was performed using ONT R9.4 flow cells and the
294 standard MiniKNOW protocol.

295
296 PacBio Sequel II sequencing was performed using 500 ng of CapTrap samples following the
297 SMRTbell™ Express Template Prep Kit 2.0 protocol.

298

299 **R2C2 preparation for ONT sequencing of human and mouse**

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

322

323 **cDNA preparation for ONT sequencing of human and mouse**

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

332 amplified using cDNA primers and LongAmp Hot Start Master Mix (NEB, M0533S). Samples
333 were treated with NEB exonuclease I (NEB, M0293) for 15 mins at 37⁰C to degrade linear
334 single-stranded DNA, followed by enzyme inactivation at 80⁰C for 15 mins. Samples were
335 purified with Agencourt AMPure XP beads. Elution was done with 12 ul of elution buffer. 1ul of
336 libraries was electrophoresed on TapeStation screentapes to assess size distribution, quantity
337 and quality of library. FLO-MIN106D flow cells were primed with EXP-FLP002 kit reagents
338 followed by loading of PCR-cDNA library mixed with rapid adapter F (along with sequencing
339 buffer and loading beads). Sequencing of the library was performed without any size selection
340 using MinION Mk1B devices and MinKNOW software interface.
341

342 **dRNA preparation for ONT sequencing of human and mouse**

343 dRNA libraries were prepared from 75ug total RNA. RNA samples were poly-A selected using
344 the NEXTFLEX poly-A kit. Purified mRNA was eluted in 12uL NF H₂O. Library preparation was
345 performed on purified mRNA using the SQK-RNA002 kit. Direct RNA RT adapters were
346 annealed and ligated to full-length mRNA using T4 DNA Ligase, NEBNext Quick Ligation
347 Reaction Buffer, and Nanopore's RNA CS. Adapter-ligated mRNA was incubated with dNTPs,
348 5x first-strand buffer, nuclease-free water, SuperScript IV, and 0.1M DTT to create a cDNA-RNA
349 hybrid. This reverse-transcription (RT) step is recommended by Nanopore to reduce secondary
350 structure formation of the mRNA as it is being sequenced. RTed RNA was purified using
351 RNAClean XP beads. Nanopore adapters were ligated onto the RTed RNA using NEBNext
352 Quick Ligation Reaction Buffer and T4 DNA Ligase. Following RNAClean XP bead cleanup, the
353 libraries were eluted in 21uL of Nanopore's Elution Buffer. 1 uL of each library was quantified on
354 the TapeStation to ensure nucleic acid concentration was at minimum ~200ng. Libraries were
355 loaded into MinION flow cells using the EXP-FLP002 Flow Cell Priming Kit. Libraries were
356 sequenced for 72 hour runs.
357

358 **Manatee ONT genome sequencing**

359 2 µg of genomic DNA in a total volume of 100 µl was fragmented with the g-Tube fragmentation
360 method (Covaris, Woburn, MA, USA) by using centrifugation at 6,000x g for 1 min. The large
361 DNA fragments were enriched by using 0.85x volume of Agencourt AMPure XP beads
362 (Beckman Coulter, Brea, CA, USA) in the purification procedure. The enriched DNA fragments
363 were subjected to library preparation with Nanopore Genomic DNA Ligation Sequencing Kit

364 (Oxford Nanopore Technologies, Oxford, UK) following the manufacture's protocol. A total of
365 700 ng of final library product was loaded on a flow cell and sequenced with a Nanopore
366 GridION sequencer (Oxford Nanopore Technologies, Oxford, UK) for a 72-hr run. A total of 5
367 flow-cell runs were conducted for this project.

368 **Manatee cDNA Pacbio library preparation and sequencing**

369 Approximately 280 ng of total pooled RNA were processed according to a modified IsoSeq
370 protocol. The sample was spiked-in with the uncapped E2 RNA variant control mix (SIRVs,
371 Lexogen, Cat # 025.03) at a 2.83% mass proportion relative to the total RNA. The resulting
372 mixture was subjected to a globin removal step using the QIAseq FastSelect™- HRM Globin
373 removal reagent (cat # 334376). This kit was designed for globin removal from human, mouse,
374 and rat tissues and was found to perform with various degrees of efficiency on blood from a
375 wide variety of samples of mammalian origin. Globin removal was performed as recommended
376 in the QIAseq FastSelect™- rRNA HRM -Globin Handbook (Oct 2019) in the NEBNext Ultra II
377 section, except that the high-temperature fragmentation step was omitted. The globin removal
378 reaction (9 µl) contained: 280 ng sample (RNA plus 2.83% SIRVs), QIAseq FastSelect globin
379 removal reagent, 2 µl NEBNext Single Cell RT Primer Mix (NEB #6421), and 2.25 µl of
380 NEBNext Single Cell RT buffer (4x). This mixture was prepared in a 0.2 ml PCR tube and
381 subjected to a stepwise series of 2 min incubations each of 75°C, 70°C, 65°C, 60°C, 55°C, 37°C
382 and 25°C. At this point, the sample was snap-cooled by transferring to a pre-chilled freezer
383 block until ready for the RT and amplification steps. From this point on, cDNA synthesis was
384 done as described in the "Protocol for Low Input RNA: cDNA Synthesis and Amplification" (NEB
385 #E6421) starting on section 2.3. More specifically, the template "RT and Template Switching"
386 reaction consisted of 9 µl of globin-removed RNA, 2.75 µl NEBNext Single Cell RT Buffer (4x), 1
387 µl of NEBNext Template Switching Oligo, 2 µl of NEBNext Single Cell RT Enzyme Mix and
388 enough water to bring the total to 20 µl. The reaction was incubated in a thermocycler for 90 min
389 at 42 °C and 10 min at 72 °C. The cDNA products were split into four aliquots for PCR
390 amplification (100 µl) reactions containing 2 µl NEBNext Single Cell cDNA PCR Primer, 0.5 µl
391 10X NEBNext Cell Lysis Buffer, 50 µl NEBNext Single Cell cDNA PCR Master Mix, 5 µl RT and
392 Template Switching reaction and water. Amplified cDNA was purified by AMPure, one round at
393 0.8 to 1.0 beads to sample ratio and one round at 0.65:1.0 ratio. The yield of amplified cDNA by
394 this modified protocol (300-400 ng) was about 10-fold lower than the standard protocol (i.e.,
395 without globin-removal). The average cDNA size was ~1400 bp. When increased amounts of
396 cDNA were desired the cDNA was amplified by 5 additional PCR cycles.

397 Two preps obtained with the above described protocol were pooled together and 500 ng were
398 loaded on an electrophoretic lateral fractionation system (ELF, SageScience). Fragments above
399 2.5 kb were collected, re-amplified (10 cycles), and re-pooled equimolarly with non-size-
400 selected cDNA fragments. This re-pooled cDNA prep is referred to as “enriched cDNA_>2.5kb”.
401 Both non_enriched cDNA and enriched cDNA_>2.5kb cDNA were used for SMRT bell library
402 was constructed starting with 1 µg of cDNA as described (PacBio IsoSeq protocol 101-070-200
403 Version 06, September 2018). Briefly, SMRTbell adaptors (Iso-Seq™) were added using
404 reagents from the PacBio SMRTbell Template Prep Kit 1.0-SPv3 starting with either 200 ng (for
405 enriched cDNA >2.5kb) or 700 ng (for non enriched cDNA). The main steps included: DNA
406 Damage Repair, End Repair, Blunt-end ligation of SMRT bell adaptors, and ExoIII/ExoVII
407 treatment. This procedure resulted in ~25-30% yield. Finally, libraries were eluted in 15 ul of 10
408 nM Tris HCl, pH 8.0. Library fragment size was estimated by the Agilent TapeStation (genomic
409 DNA tapes), and this data was used for calculating molar concentrations.
410 The enriched cDNA >2.5 kb library was diffusion-loaded on a single SEQUEL SMRT cell
411 (University of Florida, ICBR-NGS core lab) at loading concentration was 10 pM, using 4-hr pre-
412 extension, 20 hr movies and v3 chemistry reagents (for binding and sequencing). All other steps
413 for sequencing were done according to the recommended protocol by the PacBio SMRT Link
414 Sample Setup and Run Design modules (SMRT Link 6.0).

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

417 **Manatee cDNA Nanopore library preparation and sequencing**

418 One hundred and fifty nanograms of total pooled RNA were processed according to a modified
419 ONT cDNA-PCR Sequencing protocol (cDNA-PCR-PCS109, version PCS_9085 v109 revJ Aug
420 14, 2019). Spike-in and globin depletion treatment was conducted as described for Pacbio library
421 preparation. In this case, the globin removal reaction (11 ul) contained: sample (RNA plus SIRVs),
422 globin removal reagent, 1 mM dNTP, 0.2 µM VPN primer from the Nanopore cDNA synthesis
423 protocol (i.e., in place of random primers), and 1X RT buffer (ThermoFisher). This mixture was
424 prepared in a 0.2 ml PCR tube and submitted to a stepwise series of 2 min incubation for each of
425 75 °C, 70 °C, 65 °C, 60 °C, 55 °C, 37 °C and 25 °C. At this point, the sample was snap-cooled by
426 transferring to a pre-chilled freezer block until ready for the RT and amplification steps. From this
427 point on, cDNA synthesis was done as described in the cDNA-PCR Sequencing (SQK-PCS109)
428 Nanopore manual starting on page 9 (Version: PCS_90985_v109_revJ_14Aug2019). A single
429 globin removal and cDNA synthesis reaction was split into four PCR reactions for amplification.

430 This process resulted in approximately 2 micrograms of “full-length” cDNA with an average size
431 of ~1800 bp. One size-selected library was constructed by loading 1500 ng of this cDNA on an
432 electrophoretic lateral fractionation system (ELF, SageScience), collecting 5 kb were collected,
433 re-amplifying (6 cycles) and re-pooling with non-size-selected cDNA fragments. Adaptor ligation
434 and sequencing were performed according to the cDNA-PCR Sequencing (SQK-PCS109)
435 Nanopore manual. Between 120-140 fmol of cDNA was loaded on a FLO-MIN106D (R9.4
436 SpotON) flow cell for sequencing on the minION device. Two runs were done on non-size-
437 selected manatee cDNA, while only one run was done on the cDNA that had been enriched with
438 >2.5 kb fragments. Sequencing runs were allowed to proceed for 48 hours.
439

440 **Long-read data processing**

441 Basecalling of ONT data from human, mouse and manatee was performed with Guppy 4.2.2
442 and hac 9.4.1 config file, with default parameters, except: --qscore_filtering --min_qscore 7
443 (these non-default parameters were used in all ONT cDNA runs except for R2C2 datasets).
444 Direct RNA basecalling was also performed with Guppy 4.4.2 with the following configurations: -
445 -qscore_filtering yes --min_qscore 7 --reverse_sequence yes
446 --u_substitution yes
447

448 PacBio full-length non-chimeric (FLNC) reads were generated with CCS 4.2.0 (parameters: --
449 noPolish --minLength=10 --minPasses=3 --min-rq=0.9 --min-snr=2.5), Lima 1.11.0 (parameters:
450 FASTA with the appropriate adapters --isoseq --min-score 0 --min-end-score 0 --min-signal-
451 increase 10 --min-score-lead 0), and Refine 3.3.0 (parameters: --min-polya-length 20 --require-
452 polya).
453

454 Consensus R2C2 reads were generated with C3POa v1.0.0
455 (<https://github.com/rvolden/C3POa/tree/gonk>) with default options
456

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

460 **Reference genome and annotations**

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

468

469 **Simulated data**

470 Simulating RNA reads simply from the reference transcriptome would only allow the
471 assessment reconstruction of known transcript models. Thus, we extended both human and
472 mouse annotations with artificial novel transcripts. To obtain those, we mapped reference
473 transcripts of an undisclosed mammalian organism to the human and mouse genomes and
474 converted the alignments into transcript models using SQANTI²⁰. We then arbitrarily selected
475 isoforms of known genes that have only canonical splice sites (GT-AG, GC-AG and AT-AC) and
476 merged them into human and mouse GENCODE Basic annotations.

477

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

482

483 To simulate reads produced by different sequencing platforms we used existing simulation
484 methods. Illumina 2x150bp read pairs were generated with the RSEM simulator²¹ using an
485 error model obtained from real RNA-Seq data²² (accession number ERR1474891). ONT reads
486 were simulated with NanoSim²³ using pre-trained cDNA and dRNA models available in the
487 package with average error rate of 15.9% (4.8% substitutions, 6.0% deletions, 5.1% insertions)
488 and 11.2% (2.8% substitutions, 5.9% deletions, 2.5% insertions) respectively. PacBio CCS
489 reads were obtained with IsoSeqSim (<https://github.com/yunhaowang/IsoSeqSim>), which
490 truncates input reference transcript sequences and uniformly inserts errors according to given
491 probabilities. We used Sequel II truncation probabilities provided along with the package. Error
492 rate was estimated using real PacBio cDNA CCS reads obtained in this work as 1.6% (0.4%

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

495

496 We simulated two datasets containing reads from all 3 platforms listed above but with slightly
497 different properties. Human datasets were simulated with 100 million Illumina read pairs, 30
498 million ONT cDNA and 10 million PacBio reads. Mouse datasets also contained 100 million
499 Illumina read pairs, but equal amounts of PacBio CCS and ONT dRNA reads were generated
500 (20 million sequences each).

501

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

505

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

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

511

512 The 15 µg of WTC-11 RNAs from each biological replicate, ENCODE BioSample Accession
513 #ENCBS944CBA and #ENCBS474NOC, were used for the single strand (ss)CAGE library
514 preparation described in the published protocol²⁴. Briefly, the 15 µg RNAs were aliquoted to 5
515 µg in three tubes and reverse transcribed to cDNAs with random primers, and the RNA-cDNA
516 hybrids were cap-trapped by the streptavidin beads. The single strand cDNAs were released
517 from the beads and ligated to the Illumina adaptors with an index. 1080 amols of the cap-
518 trapped single strand cDNAs from each biological replicate were sequenced by Illumina HiSeq
519 Rapid SBS Kits v2 (SR, 150 cycles, 1 lane for each), producing approximately 40 million reads
520 per sample.

521

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

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

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

527

528 **Full-length transcript validation with NRSeq**

529 Depending on sample availability, we may further sequence the WTC-11 cell line using NRSeq,
530 a method for direct RNA sequencing which can distinguish full length reads (i.e., dRNA reads
531 containing both the 5' cap and polyA tail)²⁵. NRSeq uses an oligomer adaptation approach to
532 ligate an adapter specifically to 5' m7G capped RNAs and performs polyA-selected direct RNA
533 sequencing. NRSeq would provide additional data to validate start-to-end RNA transcript
534 sequence without RT-PCR artifacts²⁶. Because the technique requires approximately 2.5 ug of
535 poly(A)-selected RNA, the sequencing will be performed on two independent biological
536 replicates of WTC-11 that were not from the original cell batch from which long-read sequencing
537 was performed.

538

539 **GENCODE benchmarks and computational evaluation**

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

549

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

555

556 Annotators will also take advantage of local alignment tools integrated into annotation software
557 to give further alternative views of alignments and improve annotation accuracy. Transcript
558 models will be manually extrapolated from the alignments by annotators using the otter
559 annotation interface²⁸. Alignments will be navigated using the Blixem alignment viewer^{29,30} and
560 where required visual inspection of the dot-plot output from the Dotter tool³¹ will be used to
561 resolve any alignment with the genomic sequence that was unclear or absent from Blixem.
562 Short alignments (<15 bases) that cannot be visualized using Dotter will be detected using
563 Zmap DNA Search³¹ (essentially a pattern matching tool). The construction of exon-intron
564 boundaries will require the presence of canonical splice sites (defined as GT-AG, GC-AG and
565 AT-AC) and any deviations from this rule will be given clear explanatory tags (for example non-
566 canonical splice site supported by evolutionary conservation). All non-redundant splicing
567 transcripts at an individual locus will be used to build transcript models, and all alternatively
568 spliced transcripts will be assigned an individual biotype based on their putative functional
569 potential. Once the correct transcript structure has been ascertained the protein-coding potential
570 of the transcript will be determined on the basis of its context within the locus, similarity to
571 known protein sequences, the sequences of orthologous and paralogous proteins, candidate
572 coding regions (CCRs) identified by PhyloCSF, evidence of translation from mass spectrometry
573 and Ribo-seq data, the presence of Pfam functional domains, the presence of possible
574 alternative ORFs, the presence of retained intronic sequence and the likely susceptibility of the
575 transcript to nonsense-mediated mRNA decay (NMD). Although the annotation of transcript
576 functional biotype and CDS is not required of submitters, it will be added to transcripts as a
577 matter of routine manual annotation and may be used to investigate the detection or non-
578 detection of groups of transcripts by submitters. Where necessary, annotations will be checked
579 by a second annotator to ensure completeness and consistency of annotation between the
580 genes annotated for LRGASP and the remainder of the Ensembl/GENCODE geneset.

581 **Computational evaluation of transcript isoform detection and quantification**

582 *Challenge 1 Evaluation: Transcript isoform detection*

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

- 585 1. Lexogen SIRV-Set 4 (SIRV-Set 3 plus 15 new long SIRVs with sizes ranging from 4 to
586 12 kb)

- 587 2. Comprehensive GENCODE annotation: human v39, mouse vM28. GENCODE human
588 v28 and vM27 are available at the time of the LRGASP data release and new versions of
589 GENCODE will be released after the close of LRGASP submissions.
- 590 3. A set of transcripts from a subset of undisclosed genes which will be manually curated
591 by GENCODE. These transcripts will thus be considered high-quality models derived
592 from LRGASP data
- 593 4. Simulated data for both Nanopore (Nanosim) and PacBio (Iso-SeqSim) reads
594

595 The rationale for including these different types of transcript data is that each set creates a
596 different evaluation opportunity, but also has its particular limitations. For example, SIRVs and
597 simulated data provide a clear ground truth that allows the calculation of standard performance
598 metrics such as sensitivity, precision or false discovery rate. Evaluation of SIRVs can identify
599 potential limitations of both library preparation as well as sequencing, but the SIRVs themselves
600 represent a dataset of limited complexity. Higher complexity can be generated when simulating
601 long reads based on actual sample data. However, read simulation algorithms only capture some
602 potential biases of the sequencing technologies (e.g., error profiles) and not of the library
603 preparation protocols. In any case, both types of data approximate, but do not fully recapitulate
604 real-world datasets. Evaluation against the GENCODE annotation¹⁵ represents this real dataset
605 scenario, although in this case the ground truth is not entirely known. This limitation will be partially
606 mitigated by the identification of a subset of GENCODE transcript models that will be revised and
607 deemed as high-confidence by GENCODE curators, and by follow-up experimental validation for
608 a small set of transcripts using semi-quantitative RT-PCR and quantitative PCR (qPCR)
609 approaches. In this way, although an exhaustive validation of the real data is not possible,
610 estimates of the methods' performances can be inferred. By putting together evaluation results
611 obtained with all these different benchmarking datasets, insights will be gained on the
612 performance of the library preparation, sequencing and analysis approaches both in absolute and
613 in relative terms.

614
615 The evaluation of the transcript models will be guided by the use of SQANTI categories²⁰ (**Fig**
616 **2a**), implemented in the SQANTI3 software (<https://github.com/ConesaLab/SQANTI3>), and will
617 incorporate additional definitions and performance metrics to provide a comprehensive
618 framework for transcript model assessment (**Table 2**). The evaluation considers the accuracy of
619 the transcript models both at splice junctions and at 3'/ 5' transcript ends. It will take into
620 account external sources of evidence such as CAGE data, polyA annotation and support by

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

625

626

627 **Table 2: Transcript Classifications and Definitions used by the LRGASP computational**
 628 **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)

630

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

632

633 *SIRVs*

634 In order to evaluate SIRVs, we will extract from each submission all transcript models that
635 associate to SIRV sequences after SQANTI3 analysis. This not only includes FSM and ISM
636 isoforms of SIRVs, but also NIC, NNC, antisense and fusion transcripts mapping to SIRV loci.

637 The metrics for SIRV evaluation are defined as follows.

638

639 **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)

640

641

642 *Simulated Data*

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

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

649

650 **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

651

652

653 *Comprehensive GENCODE annotation*

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

660

661 **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

662

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

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

668

669

670 *Challenge 2 Evaluation: Transcript isoform quantification*

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

678

679 *Evaluation metrics*

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

681 1) Single sample data when the ground truth is available

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

683 3) Multiple replicates when ground truth is not available

684

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

688

689 Single sample data (ground truth is available)

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

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

695

696 •Spearman Correlation Coefficient (SCC)

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

700

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

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

704 •Abundance Recovery Rate (ARR)

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

706

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

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

708

709 •Median Relative Difference (MRD)

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

712

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

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

714

715 •Normalized Root Mean Square Error (*NRMSE*)

716 *NRMSE* provides a measure of the extent to which the one-to-one relationship deviates from a

717 linear pattern. It can be calculated by

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

718

719 where s_{Θ} is the sample standard deviation of Θ .

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

721

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

723

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

725 Denote $\hat{\theta}_{ijk}$ and θ_{ijk} as the estimation and ground truth of transcript isoform i ($i = 1, 2, \dots, I$) in

726 a sample, where j ($j = 1, 2$) represents different groups (i.e., conditions or tissues) and

727 k ($k = 1, 2, \dots, K$) represents different replicates within the group j .

728

729 We assess the quantification performance by ROC (receiver operating characteristic) analysis

730 of identifying true differentially expressed transcript isoforms. At first, we define Average Log

731 Fold Change (*ALFC*) of transcript isoform i as:

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

732

733 Next, based on the ground truth values and a given threshold (e.g., 1 as below), we can define

734 whether a transcript isoform is truly differentially expressed or not:

735 Positives (truly differentially expressed)

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

737 Negatives (not truly differentially expressed)

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

738

739 Based on the estimated values, we can also obtain the “predicted positives” and “predicted

740 negatives” with the same threshold. Therefore, we can identify “true positives”, “true negatives”,

741 “false positives” and “false negatives” to calculate the ROC-based statistics, including precision,

742 recall, accuracy, F1-score, AUC and pAUC, and also plot ROC (**Supplementary Fig. S2**).

743

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

748

749 Multiple replicates under different conditions (without the ground truth)

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

754

755 •Reproducibility

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

758
$$RM = \sqrt{\frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J s_{ij}^2}$$

759 Here, s_{ij} is the sample standard deviation of $\log(\hat{\theta}_{ijk} + 1)$ ($k = 1, 2, \dots, K$), which is
760 calculated by

761
$$s_{ij} = \sqrt{\frac{1}{K} \sum_{k=1}^K \left(\log(\hat{\theta}_{ijk} + 1) - u_{ij} \right)^2},$$

762 where

763
$$u_{ij} = \frac{1}{K} \sum_{k=1}^K \log(\hat{\theta}_{ijk} + 1).$$

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

767

768 •Consistency

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

$$773 C(\alpha) = \frac{1}{IJ \cdot C_K^2} \sum_{i=1}^I \sum_{j=1}^J \sum_{1 \leq k_1 < k_2 \leq K} P\left(\left\{\log(\hat{\theta}_{ijk_1} + 1) < \alpha, \log(\hat{\theta}_{ijk_2} + 1) < \alpha\right\} \text{ or } \left\{\log(\hat{\theta}_{ijk_1} + 1) \geq \alpha, \log(\hat{\theta}_{ijk_2} + 1) \geq \alpha\right\}\right),$$

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

775

776 •Resolution Entropy (*RE*)

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

$$780 RE = - \sum_{m=1}^M P_m \ln(P_m), \text{ where } P_m = \frac{n_m}{\sum_{j=1}^M n_j}.$$

781 Here, the abundance estimates are binned into M groups, where n_m represents the number of
 782 transcript isoforms with the abundance estimate $\hat{\Theta} \in [m \cdot \alpha, (m + 1) \cdot \alpha)$, and

783 $\alpha = \max(\hat{\Theta}) / M$. $RE = 0$ if all transcript isoforms have the same estimated abundance

784 values, while it obtains a large value when the estimates are uniformly distributed among M

785 groups.

786

787 *Evaluation with respect to multiple transcript features*

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

793

794 • K-value

795 Most methods for transcript isoform quantification assign sequencing coverage to isoforms;

796 therefore, the exon-isoform structure of a gene is a key factor influencing quantification

797 accuracy. Here, we use a statistic K-value (manuscript in preparation, **Supplementary Fig. S6**)

798 to measure the complexity of exon-isoform structures for each gene. Suppose a gene of interest

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

$$801 \quad a_{ie} = \begin{cases} 1, & \text{if the isoform } i \text{ includes the exon } e \\ 0, & \text{otherwise} \end{cases}$$

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

$$803 \quad \text{K-value} = \frac{\sigma_{max}(A)}{\sigma_{min}(A)},$$

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

806

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

813

814

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

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

822

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

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

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

849

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

853

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

857

858 We will compare metric statistics across analysis pipelines and sequencing platforms. A number
859 of genes will be selected for PCR-based experimental validation (see below), including
860 transcripts of cytokine genes, which have been studied by LRGASP consortium members in
861 detail³². Ferrante et al.³² designed and validated primers to measure cytokine transcript levels in
862 Florida manatees from blood samples, specifically for interleukin (IL)-2, -6, -10, interferon-

863 gamma (INF-gamma) and Tumor necrosis-alpha (TNF-alpha), and these methods will be
864 adopted for development of isoform-specific assays.
865

866 **Experimental validation of transcript models and expression estimates**

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

873 *Challenge 1 Evaluation: Transcript isoform detection*

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

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

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

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

896 We plan to select 96 targets from human WTC-11 cells and 96 targets from the mouse
897 129/Castaneus cells. Each target will comprise a sequence region 300 to 1500 bp long. Two
898 replicates each from the WTC-11 and 129/Castaneus sample will be apportioned for a reverse-
899 transcriptase reaction followed by target amplification using isoform-specific primers. We will
900 conduct the assay in plate format to allow for high-throughput processing. All products following
901 RT-PCR will be pooled and subjected to long-read sequencing for validation. A subset of these
902 samples will be selected for Sanger sequencing. Table 6 shows the breakdown of targets we
903 will select.

904

Category	WTC-11 (Human)	129/Castaneus (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

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

907

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

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

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

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

924 In addition to the validation using a PCR-based approach (Table 6), high-throughput validation
925 of full-length transcripts will be obtained by application of the NRSeq strategy²⁵ on WTC-11
926 cells, which does not rely on PCR. NRSeq employs a chemical labeling strategy to add a
927 signature oligonucleotide exclusively to the 5' caps of mRNAs, thus, full-length mRNA
928 sequences from the 5' cap to the polyA sequence may be distinguished from incomplete
929 sequence fragments. We will compare NRSeq data generated WTC-11 against models
930 submitted by participants.

931 *Challenge 2 Evaluation: Transcript isoform quantification*

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

939

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

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

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

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

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

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

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

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

968 **Challenge submissions and timeline**

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

971

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

- 974 ● Challenge 1: transcript isoform detection with a high-quality genome (iso_detect_ref)
- 975 ○ Samples
- 976 ■ WTC11 (human iPSC cell line)
- 977 ■ H1_mix (human H1 ES cell line mixed with human Definitive Endoderm
- 978 derived from H1)
- 979 ■ ES (mouse ES cell line)
- 980 ■ human_simulation - simulated human reads (Illumina, ONT, and PacBio
- 981 cDNA)
- 982 ■ mouse_simulation - simulated mouse reads (Illumina and PacBio cDNA,
- 983 ONT dRNA)
- 984 ○ Result files:
- 985 ■ models.gtf.gz
- 986 ■ read_model_map.tsv.gz
- 987 ● Challenge 2: transcript isoform quantification (iso_quant)
- 988 ○ Samples
- 989 ■ WTC11 (human iPSC cell line)
- 990 ■ H1_mix (human H1 ES cell line mixed with human Definitive Endoderm
- 991 derived from H1)
- 992 ■ human_simulation - simulated human reads (Illumina, ONT, and PacBio
- 993 cDNA)
- 994 ■ mouse_simulation - simulated mouse reads (Illumina and PacBio cDNA,
- 995 ONT dRNA)
- 996 ○ Result files:
- 997 ■ expression.tsv.gz
- 998 ■ models.gtf.gz
- 999 ● Challenge 3: de novo transcript isoform detection (iso_detect_de_novo)
- 1000 ○ Samples
- 1001 ■ Manatee (manatee whole blood)
- 1002 ■ ES (mouse ES cell line)
- 1003 ○ Result files:
- 1004 ■ rna.fasta.gz
- 1005 ■ read_model_map.tsv.gz

1006 A submission to a challenge is an entry, consisting of one or more experiments. Each entry
1007 must meet the following requirements:

1008

1009 *Requirements for Challenge 1 and 2*

1010 At least one experiment must be supplied for each sample available for a given challenge.

1011 Human and mouse samples will have biological replicates that should be used for the entry.

1012

1013 A major goal of LRGASP is to assess the capabilities of long-read sequencing for transcriptome
1014 analysis and also how much improvement there is over short-read methods. Additionally, long-
1015 read computational pipelines vary in their use of only long-read data or if they incorporate
1016 additional data for transcript analysis. To facilitate comparisons between long-read and short-
1017 read methods and variation in tool parameters, we break down submissions into different
1018 categories:

- 1019 • long-only - Use only LGRASP-provided long-read RNA-Seq data from a single sample,
1020 library preparation method and sequencing platform.
- 1021 • short-only - Use only LGRASP-provided short-read Illumina RNA-Seq data from a single
1022 sample. This is to compare with long-read approaches
- 1023 • long and short - Use only LGRASP-provided long-read and short-read RNA-Seq data
1024 from a single long-read library preparation method and the Illumina platform. Additional
1025 accessioned data in public genomics data repositories can also be used.
- 1026 • kitchen sink - Any combination of at least one LRGASP data set as well as any other
1027 accessioned data in public genomics data repositories. For example, multiple library
1028 methods can be combined (e.g. PacBio cDNA + PacBio CapTrap, ONT cDNA + ONT
1029 CapTrap+ ONT R2C2+ ONT dRNA, all data, etc.).

1030

1031 In all the above categories, the genome and transcriptome references specified by LRGASP
1032 should be used. For the long and short and kitchen sink category, additional transcriptome
1033 references can be used.

1034

1035 All replicates must be used in each experiment. Challenge 2 must report replicates separately in
1036 the expression matrix. Each team can only submit one entry per category.

1037

1038 For Challenge 1, the submitted GTF file should only contain transcripts that have been assigned
1039 a read. For Challenge 2, submitters have the option of quantifying against the reference

1040 transcriptome or a transcriptome derived from the data (i.e., results from Challenge 1). The GTF
1041 used for quantification is included as part of the Challenge 2 submission.

1042
1043 The type of platform and libraries preparation method used in a given experiment, except for
1044 kitchen sink experiments, is limited to data from a single library preparation method plus
1045 sequencing technology (long-only). LRGASP Illumina short-read data of the same sample may
1046 optionally be used in an experiment with the LRGASP long-read data (long and short)

- 1047 ● Illumina cDNA - short-only
- 1048 ● Pacbio cDNA - long-only or long and short
- 1049 ● Pacbio CapTrap - long-only or long and short
- 1050 ● ONT cDNA - long-only or long and short
- 1051 ● ONT CapTrap - long-only or long and short
- 1052 ● ONT R2C2 - long-only or long and short
- 1053 ● ONT dRNA - long-only or long and short

1054

1055 *Requirements for Challenge 3*

1056 At least one experiment must be supplied for each sample available for the challenge. Mouse
1057 samples will have biological replicates that should be used for the entry.

1058 For similar reasons as described above, the data used for a given experiment must fit in one of
1059 the following categories:

- 1060 ● long-only - Use only LGRASP-provided long-read RNA-Seq data from a single sample,
1061 library preparation method and sequencing platform. No genome reference can be used.
- 1062 ● short-only - Use only LGRASP-provided short-read Illumina RNA-Seq data from a single
1063 sample. This is to compare with long-read approaches. No genome reference can be
1064 used.
- 1065 ● long and short - Use only LGRASP-provided long-read and short-read RNA-Seq data
1066 from a single long-read library preparation method and the Illumina platform. No genome
1067 reference can be used.
- 1068 ● long and genome - Use only LGRASP-provided long-read RNA-Seq data from a single
1069 long-read library preparation method. A genome reference sequence can be used.
- 1070 ● kitchen sink - Any combination of at least one LRGASP data set as well as any other
1071 accessioned data in public genomics data repositories. For example, multiple library
1072 methods can be combined (e.g. PacBio cDNA + PacBio CapTrap, ONT cDNA + ONT
1073 CapTrap+ ONT R2C2+ ONT dRNA, all data, etc.).

1074
1075 In all the above categories, except for kitchen sink a transcriptome reference cannot be used.
1076 The submitted FASTA file should only contain transcripts that have been assigned a read.
1077 Each team can only submit one entry per category.
1078
1079 LRGASP biological data is currently available at the ENCODE DCC
1080 (https://www.encodeproject.org/search/?type=Experiment&internal_tags=LRGASP). The
1081 simulated data is available from Synapse (<https://www.synapse.org/#!/Synapse:syn25683370>).
1082 The competition launched on May 1, 2021 and challenge submissions are expected to close on
1083 October 1, 2021.

1084

1085 **LRGASP Data QC**

1086 Initial quality control (QC) metrics were determined for the LRGASP data (**Figure 4**). Reads
1087 (ONT cDNA, dRNA, CapTrap) or consensus reads (PacBio cDNA and CapTrap and ONT
1088 R2C2) were aligned to the human or mouse genome as appropriate using minimap2 with the
1089 following parameters: -ax splice --secondary=no -G 400k. For each data type, the reads and
1090 their resulting alignments in sam format were parsed for the following parameters:

1091 1) Number of aligned reads

1092 2) Number of aligned reads with adapters on both ends

1093 For ONT dRNA this is not applicable as this workflow does not attach an adapter
1094 to the 5' end of molecules. For ONT cDNA and CapTrap this percentage was
1095 determined by pyChopper. For all other data types, all provided reads are
1096 assumed to have adapters on both ends as the pre-processing pipelines (lima
1097 and C3POa) discard reads otherwise.

1098 3) median read length

1099 measured by the number of aligned bases (matches or mismatches)

1100 4) median accuracy

1101 measured by $\text{matches}/(\text{matches}+\text{mismatches}+\text{indels})$,

1102 5) Percent of aligned reads where the orientation of the reads as determined by 5' and 3'
1103 adapter sequences agrees with the direction of the read alignment

1104 determined by minimap2 through splice site context (calculated only for the
1105 subset of reads with splice alignments with the ts:A: flag in their sam entry),

1106 6) Percent of reads originating from spike-in molecules

1107 determined by alignment to the SIRVomeERCC fasta entry in the genome
 1108 sequence files
 1109 7) Pearson correlation between replicates
 1110 determined by quantifying gene expression for each replicate and calculating the
 1111 pearson r value based on those expression values.

1112
 1113
 1114 **Table 7: Summary statistics for LRGASP data.** For each sample, replicates were combined
 1115 when reporting statistics.

Sample	ES					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	Sequell	Sequell
# 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.						

1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126

Sample	WTC11					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	Sequelll	Sequelll
# 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.						

1127

Sample	H1_mix					
Method	dRNA	cDNA	R2C2	CapTrap	CapTrap	cDNA
Tech	ONT	ONT	ONT	ONT	PacBio	PacBio
Platform	MinION	MinION	MinION	MinION	Sequelll	Sequelll
# 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	76.15¹	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.						

1128

1129

1130

1131

1132

1133

1134

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		

1135

1136

1137 **Data and code availability**

1138 All code and documentation associated with the LRGASP Consortium can be found through
 1139 <https://www.genencodegenes.org/pages/LRGASP/> and <https://github.com/LRGASP>.

1140

1141

1142

1143

1144

1145 Acknowledgments

1146 We thank Lexogen, Oxford Nanopore Technologies (ONT), and Pacific Biosciences for helpful
1147 discussions. ONT provided partial support of flow cells and reagents. We thank Xingjie Ren and
1148 Yin Shen for providing WTC11 cells, Takayo Sasaki and Dave Gilbert for providing the F121-9
1149 hybrid mouse ES cells, and Alyssa Cousineau, Krishna Mohan Parsi, and Rene Maehr for
1150 providing human H1 and H1-DE cells. We also thank Mark Akesson and Miten Jain for providing
1151 resources and technical advice for Nanopore sequencing. We thank Julie Visser for contributing
1152 artwork that gives an overview of the LRGASP Consortium. The project is supported by the
1153 following grants: Pew Charitable Trust (A.N.B.), NIGMS R35GM138122(A.N.B.), NHGRI
1154 U41HG007234 (J.L, M.D., R.G. and S.C-S) and UM1 HG009443 (A.M. and B.W.), an
1155 institutional fund of the Department of Biomedical Informatics, The Ohio State University
1156 (K.F.A.), NHGRI R01HG008759 (K.F.A.), SPBU 73023672 (A.P). J.E.L., J.M.M. and A.F. are
1157 supported by National Human Genome Research Institute of the National Institutes of Health
1158 [U41HG007234]; the content is solely the responsibility of the authors and does not necessarily
1159 represent the official views of the National Institutes of Health; Wellcome Trust
1160 [WT108749/Z/15/Z, WT200990/Z/16/Z]; European Molecular Biology Laboratory. We
1161 acknowledge Ellie Schiller Homosassa Springs Park for providing archive Lorelei blood
1162 samples.

1163 Competing Interests

1164 Design of the project was discussed with Oxford Nanopore Technologies (ONT), Pacific
1165 Biosciences, and Lexogen. ONT provided partial support of flow cells and reagents. S.C-S and
1166 A.N.B. have received reimbursement for travel, accommodation and conference fees to speak
1167 at events organised by ONT.

1168

- 1169 1. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing.
1170 *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4821–30 (2013).
- 1171 2. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the
1172 human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
- 1173 3. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford
1174 Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100
1175 (2017).
- 1176 4. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat.*
1177 *Methods* **15**, 201–206 (2018).
- 1178 5. Byrne, A., Cole, C., Volden, R. & Vollmers, C. Realizing the potential of full-length
1179 transcriptome sequencing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190097 (2019).
- 1180 6. Oikonomopoulos, S. *et al.* Methodologies for Transcript Profiling Using Long-Read
1181 Technologies. *Front. Genet.* **11**, 606 (2020).
- 1182 7. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics*
1183 *Bioinformatics* **13**, 278–289 (2015).
- 1184 8. Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A. & Tilgner, H. U. Getting the Entire
1185 Message: Progress in Isoform Sequencing. *Front. Genet.* **10**, 709 (2019).
- 1186 9. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq
1187 data. *Nat. Methods* **10**, 1185–1191 (2013).
- 1188 10. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat.*
1189 *Methods* **10**, 1177–1184 (2013).
- 1190 11. Reese, M. G. *et al.* Genome annotation assessment in *Drosophila melanogaster*. *Genome*
1191 *Res.* **10**, 483–501 (2000).
- 1192 12. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project.
1193 *Genome Biol.* **7 Suppl 1**, S2.1–31 (2006).
- 1194 13. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the

- 1195 sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.*
1196 **115**, 9726–9731 (2018).
- 1197 14. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper.
1198 *Genomics* **37**, 327–336 (1996).
- 1199 15. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
- 1200 16. Foote, A. D. *et al.* Convergent evolution of the genomes of marine mammals. *Nat. Genet.*
1201 **47**, 272–275 (2015).
- 1202 17. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**,
1203 171–181 (2014).
- 1204 18. Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.*
1205 **303**, 19–44 (1999).
- 1206 19. Shibata, Y. *et al.* Cloning full-length, cap-trapper-selected cDNAs by using the single-strand
1207 linker ligation method. *Biotechniques* **30**, 1250–1254 (2001).
- 1208 20. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences
1209 for quality control in full-length transcriptome identification and quantification. *Genome Res.*
1210 (2018) doi:10.1101/gr.222976.117.
- 1211 21. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or
1212 without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- 1213 22. Jo, J. *et al.* Midbrain-like Organoids from Human Pluripotent Stem Cells Contain Functional
1214 Dopaminergic and Neuromelanin-Producing Neurons. *Cell Stem Cell* **19**, 248–257 (2016).
- 1215 23. Hafezqorani, S. *et al.* Trans-NanoSim characterizes and simulates nanopore RNA-
1216 sequencing data. *Gigascience* **9**, (2020).
- 1217 24. Takahashi, H., Nishiyori-Sueki, H., Ramilowski, J. A., Itoh, M. & Carninci, P. Low Quantity
1218 single strand CAGE (LQ-ssCAGE) maps regulatory enhancers and promoters.
1219 doi:10.1101/2020.08.04.231969.
- 1220 25. Mulroney, L. *et al.* Identification of high confidence human poly(A) RNA isoform scaffolds

- 1221 using nanopore sequencing. doi:10.1101/2020.11.18.389049.
- 1222 26. Schulz, L. *et al.* Direct long-read RNA sequencing identifies a subset of questionable
1223 exons likely arising from reverse transcription artifacts. *Genome Biol.* **22**, 190 (2021).
- 1224 27. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE
1225 Project. *Genome Res.* **22**, 1760–1774 (2012).
- 1226 28. Searle, S. M. J., Gilbert, J., Iyer, V. & Clamp, M. The otter annotation system. *Genome*
1227 *Res.* **14**, 963–970 (2004).
- 1228 29. Sonnhammer, E. L. & Durbin, R. A workbench for large-scale sequence homology analysis.
1229 *Comput. Appl. Biosci.* **10**, 301–307 (1994).
- 1230 30. Sonnhammer, E. L. & Durbin, R. An expert system for processing sequence homology
1231 data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 363–368 (1994).
- 1232 31. Sonnhammer, E. L. & Durbin, R. A dot-matrix program with dynamic threshold control
1233 suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–10 (1995).
- 1234 32. Ferrante, J. A., Hunter, M. E. & Wellehan, J. F. X. DEVELOPMENT AND VALIDATION OF
1235 QUANTITATIVE PCR ASSAYS TO MEASURE CYTOKINE TRANSCRIPT LEVELS IN THE
1236 FLORIDA MANATEE (*TRICHECHUS MANATUS LATIROSTRIS*). *J. Wildl. Dis.* **54**, 283–
1237 294 (2018).

1238
1239

1240

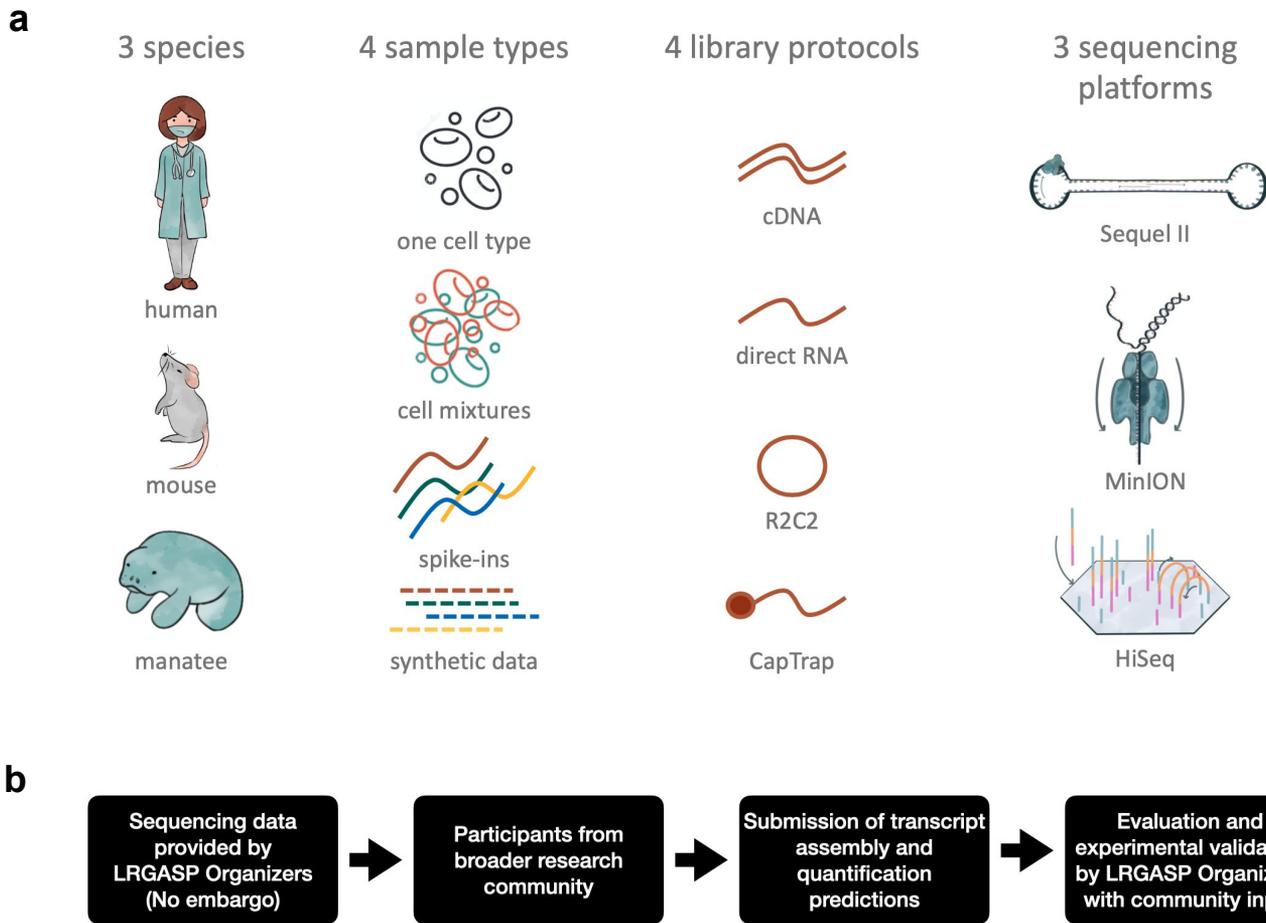


Fig. 1: Overview of the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP). a, LRGASP Consortium as a research community effort. **b,** Overview of LRGASP data.

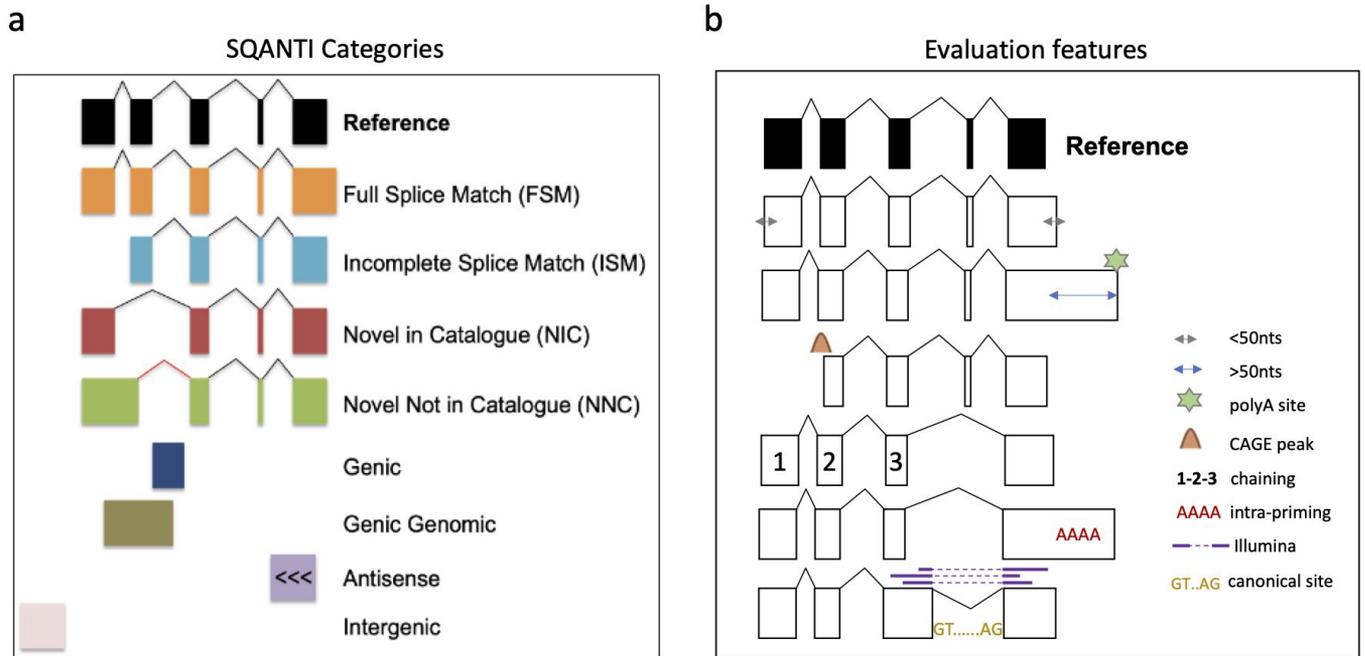


Fig. 2: SQANTI-based evaluation of transcript identification methods for Challenges 1 and 3. a, Transcripts are compared to a best matched reference transcript and categorized based on shared junctions between the reference. **b,** Additional features that are considered when evaluating transcript models

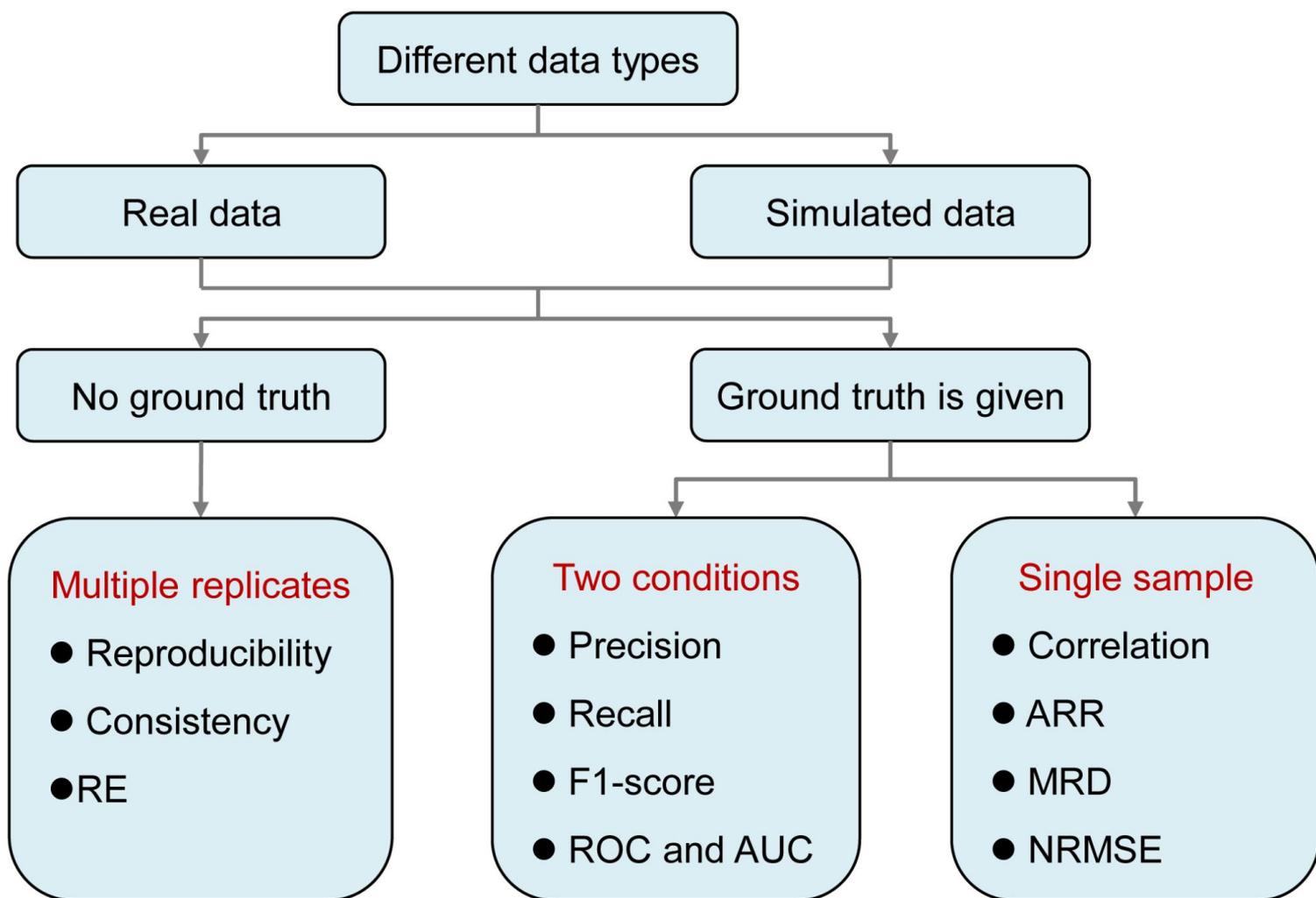


Fig. 3: Evaluation metrics of gene isoform quantification under different data types. RE - Resolution Entropy, ARR - Abundance Recovery Rate, MRD - Median Relative Difference, NRMSE - Normalized Root Mean Square Error

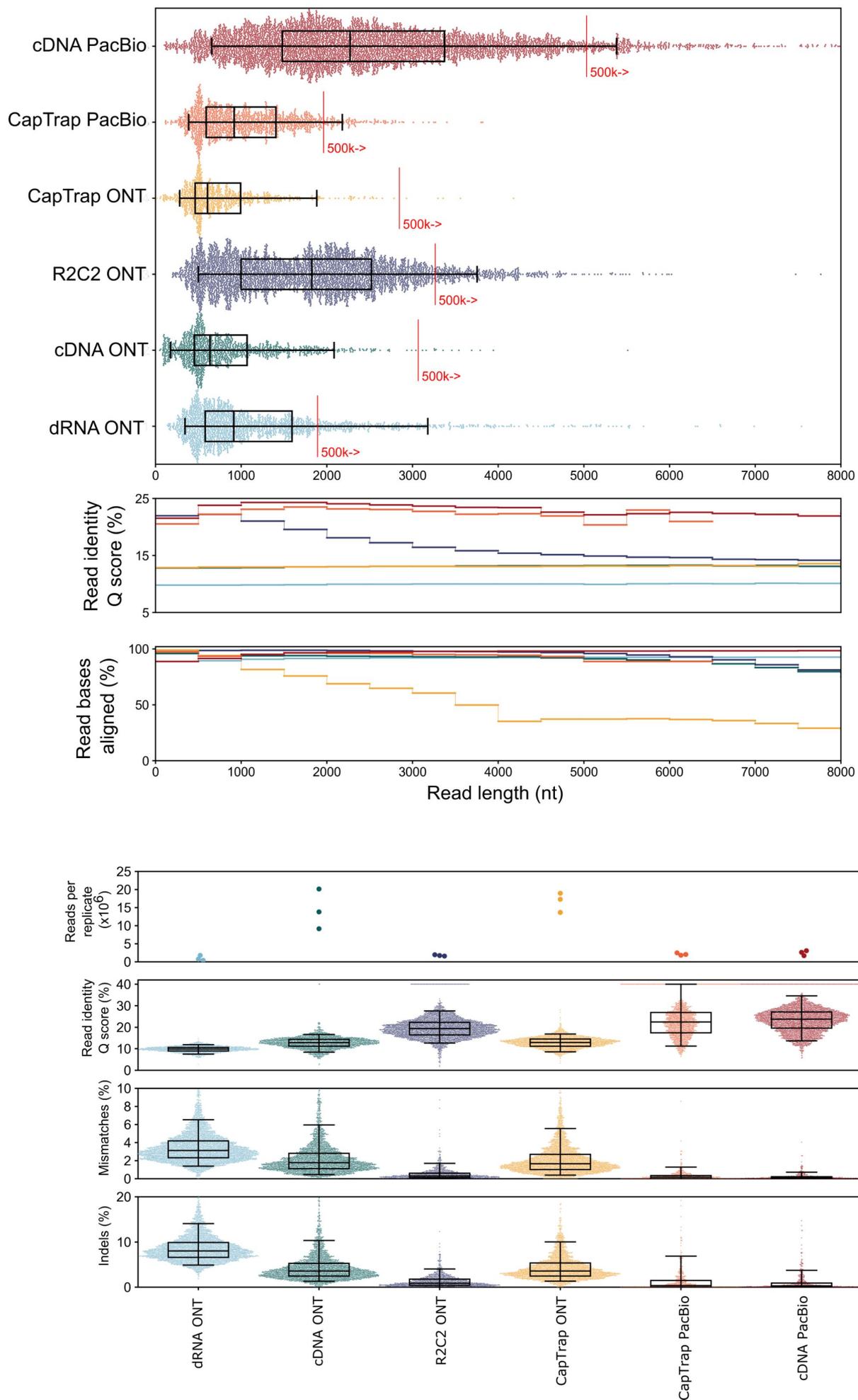


Fig. 4: Summary of LRGASP Data

Supplementary Files

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

- [SupplementaryTable1.txt](#)
- [RegisteredReportSupplInfo210730wFigures.pdf](#)
- [RegisteredReportSupplInfo210730wFigures.pdf](#)