As of September 22, 2020

Description of LRGASP Data

The LRGASP challenge encompasses different human, mouse, and manatee samples sequenced using multiple combinations of protocols and platforms. Different challenges will use distinct subsets of the samples for evaluation. The long-read sequencing platforms used in these challenges are the Pacific Biosciences (PacBio) Sequel II, Oxford Nanopore (ONT) MinION and PromethION. Samples will also be sequenced on the Illumina HiSeq 2500. The primary LRGASP library prep protocols are "standard" cDNA sequencing, direct RNA sequencing, R2C2, and CapTrap. Each sample will also include Lexogen SIRV-Set 4 spike-ins. We will also provide simulated PacBio and ONT data as part of the evaluations.

Overview of LRGASP Data. More details for each sample are described below.

Sample	# of reps	PacBio cDNA (2 SMRT cells/samp)	ONT minION cDNA	ONT minION direct RNA	R2C2 minION	CapTrap PacBio (1 SMRT cell/samp)	CapTrap minION ONT	Illumina cDNA
Human WTC-11 cell line	**3**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Mouse 129/Cast ES cell line	**3**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Human H1 ES/Def Endoderm cell line Mix	**3**	**Yes**	**Yes[#]**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Manatee Whole Blood*	**1**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	**Yes**

*SIRV Set 1 spike-ins (E2 mix); [#]ONT PromethION

Challenge 1 - Transcript isoform detection in a well annotated genome

This challenge will assess the sensitivity and precision of transcript models generated from the different platforms described above. The samples in this challenge are human iPSC and mouse ES cells that are used within the ENCODE Consortium. These were all grown and sequenced in biological triplicate using all protocols and platforms listed above.

1) **WTC11** – iPSC (Homo sapiens, derived from 30 year old Japanese male leg skin fibroblasts. At passage 41, 20/20 cells karyotyped normal. Grown by Xingjie Ren in Yin Shen's Laboratory at UCSF). 3 biological replicates:

p37 harvested on February 7, 2020: ENCODE BioSample Accession #ENCBS944CBA
p38 harvested on February 12, 2020: ENCODE BioSample Accession # ENCBS593PKA
p39 harvested on February 17, 2020: ENCODE BioSample Accession # ENCBS474NOC

2) **F1 *Mus musculus* S129/SvJae (♀) X Castaneus (♂) ES cells (F121-9)** (passage 14) (*Mus musculus*, female ES cell line, grown by Takayo Sasaki in Dave Gilbert's Laboratory at FSU Tallahassee. Description and SOP for cultivation at https://data.4dnucleome.org/biosources/4DNSRMG5APUM/

At passage 9, 20 G-banded metaphase cells were karyotyped. 18/20 showed normal female karyotype. 2/20 demonstrated non-clonal chromosome aberrations (1 cell: 40,X,add(X)(F1); 1 cell: 41,XX,+12) No evidence of trisomy 8 or 11 was detected.

Biorep#1 ENCODE BioSample Accession # ENCBS648HXY
Biorep#2 ENCODE BioSample Accession # ENCBS951CRC
Biorep#3 ENCODE BioSample Accession # ENCBS418RDP

Challenge 2 - Transcript isoforms quantification

1) **H1** – ES cells

This challenge will assess the expression levels measured by the different sequencing technologies in human. The sample used here will be a mix of human H1 and H1 differentiated into definitive endoderm (H1-DE), whose precise mixing ratio will be released after the close of the submission window. These mixtures of H1 and H1-DE were all grown separately as biological triplicates, mixed at the same ratio and sequenced in triplicates using all protocols and platforms listed above. Since the mixture contains different cellular states, the mixture is expected to contain heterogeneous transcript isoforms of the same gene at different abundances.

Participants will be comparing the expression levels of genes and transcripts in these mixtures to the expression level of genes and transcripts in WTC-11 (from Challenge 1).

Challenge 3 - *De novo* annotation in a non-model genome

This challenge will assess the performance of transcriptome annotation using the Manatee, which is about the size of the human genome. This consists of a single blood transcriptome sample sequenced as cDNA on PacBio and ONT.