## SQANTI-reads: a tool for the quality assessment of long read data in multi-sample lrRNA-seq experiments.

 $\begin{array}{c} 1 \\ 1 \end{array}$  $\begin{array}{c} 1 \\ 1 \\ 2 \end{array}$ Netanya Keil<sup>4,2</sup>, Carolina Monzó<sup>4</sup>, Lauren McIntyre<sup>4,2,3,</sup> , Ana Conesa<sup>4,</sup><br>1<br>Department of Molecular Genetics and Microbiology, University o<br>USA, 32610<br><sup>2</sup> University of Florida Genetics Institute, University of Flori  $\begin{array}{c} 1 \\ 2 \\ 3 \end{array}$ 1 | 2 3 4 | <sup>1</sup>Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, JSA, 32610<br>University of Florida Genetics Institute, University of Florida, Gainesville, FL, USA, 32610<br>UF Health Cancer Center, University of Florida, Gainesville, FL, USA, 32610<br>Institute for Integrative Systems Biology <sup>2</sup>University<br><sup>3</sup>UF Health<br><sup>4</sup>Institute fo<br>Paterna, 469 <sup>2</sup> University of Florida Genetics Institute, University of Florida, Gainesville, FL, USA, 32610 UF Health Cancer Center, University of Florida, Gainesville, FL, USA, 32610<br>Institute for Integrative Systems Biology (I2SysBio), Spanish National Research Council<br>Paterna, 46980, Spain<br>Corresponding authors: ana.conesa@cs Ĭ A Histitute for Integrative Systems Biology (I2SysBio), Spanish National Rese<br>Paterna, 46980, Spain<br>\*corresponding authors: <u>ana.conesa@csic.es</u>; mcintyre@ufl.edu

Paterna, 46980, Spain<br>
Institute Systems Biology (I2S), Spanish National Research Corresponding authors: <u>ana.conesa@csic.es;</u> mcintyre@ufl.edu Paterna, 46980, Spain<br>\*corresponding authors: ana.conesa@csic.es; mcintyre@ufl.edu  $\ddot{a}$  $\ddot{ }$ 

\*corresponding and the setting and the setting as<br>ABSTRACT<br>SQANTI-reads leverages SQANTI3, a tool for the analysis of the r  $\frac{1}{2}$  $\frac{1}{2}$  $\frac{1}{2}$ ABSTRACT<br>SQANTI-rea<br>develop a<br>number/dis  $\frac{5}{3}$ develop a quality control protocol for replicated long-read RNA-seq experiments. The<br>number/distribution of reads, as well as the number/distribution of unique junction chains<br>(transcript splicing patterns), in SQANTI3 str manter protocol replies to replies the replies to replies the replies of the number/distribution of unique junction chains<br>(transcript splicing patterns), in SQANTI3 structural categories are compiled. Multi-sample<br>visuali (transcript splicing patterns), in SQANTI3 structural categories are compiled. Multi-sample<br>visualizations of QC metrics can also be separated by experimental design factors. We<br>introduce new metrics for 1) the identificat (interrige splitting patterns), in Schoole structural categories are compiled to the visibility<br>visualizations of QC metrics can also be separated by experimental design factors. We<br>introduce new metrics for 1) the identif introduce new metrics for 1) the identification of potentially under-annotated genes and<br>putative novel transcripts and 2) variation in junction donors and acceptors. All scripts are<br>open source and customizable. Using two putative novel transcripts and 2) variation in junction donors and acceptors. All scripts are<br>open source and customizable. Using two different datasets, one from *Drosophila* and one<br>benchmark dataset from the LRGASP proj putative novel transcripts and customizable. Using two different datasets, one from *Drosophila* and one<br>benchmark dataset from the LRGASP project, we demonstrate how low coverage does not<br>automatically indicate low qualit benchmark dataset from the LRGASP project, we demonstrate how low coverage does not automatically indicate low quality and how strong/weak splicing sites can be readily identified genome wide. SQANTI-reads is open source and available for download at GitHub. genome wide. SQANTI-reads is open source and available for download at GitHub.

INTRODUCTION<br>Short-read RNA<br>studying the<br>computationally,<br>Newman et al. 2 Short-read Manusey and Commonly is the most common and constraints approach in the most dividying the transcriptome. However, in srRNA-seq, transcripts must be inferred computationally, which can lead to inaccuracies in tr studies of the transcript identification (Liu et al. 2016;<br>Studynan et al. 2018). Recent advances in single-molecule long-read sequencing technologies<br>Newman et al. 2018). Recent advances in single-molecule long-read seque Computationally, Newman et al. 2018). Recent advances in single-molecule long-read sequencing technologies<br>
have opened new avenues for transcriptome analysis (reviewed in (Marx 2023; van Dijk et al.<br>
2023)). In long-read Newman et al. 2023). In long-read RNA sequencing (IrRNA-seq), full-length transcripts can be observed as<br>2023)). In long-read RNA sequencing (IrRNA-seq), full-length transcripts can be observed as<br>single sequencing reads, have opened 2023)). In long-read RNA sequencing (IrRNA-seq), full-length transcripts can be observed as<br>single sequencing reads, allowing for direct transcript detection without the need for an<br>assembly step. However, like single sequencing reads, allowing for direct transcript detection without the need for an<br>assembly step. However, like any technology, IrRNA-seq is not without errors, and factors such<br>as mRNA degradation, library preparat single sequencing reads, and the needs allowing transcript detection without the need for an<br>assembly step. However, like any technology, IrRNA-seq is not without errors, and factors such<br>as mRNA degradation, library prepa as mRNA degradation, library preparation failures, and sequencing errors can introduce biases<br>into the data.<br>A database tracking bioinformatic tools for long-read sequencing (Amarasinghe et al. 2021)

into the data.<br>A database tracking bioinformatic tools for long-read sequencing (Amarasinghe et al. 2021)<br>identifies numerous tools for the initial processing of lrRNA-seq data to assess the accuracy of A database tracking bioinformatic tools for long-read sequencing (Amarasinghe et al. 2021)<br>identifies numerous tools for the initial processing of IrRNA-seq data to assess the accuracy of<br>base-calling and the length of the  $\frac{1}{i}$ Manham Charming diamed and the tracking cand trap and the initial processing of IrRNA-seq data to assess the accuracy of<br>base-calling and the length of the reads (e.g., pycoQC (Leger and Leonardi 2019), longQC<br>(Fukasawa et ides of the length of the reads (e.g., pycoQC (Leger and Leonardi 2019), longQC<br>(Fukasawa et al. 2020), nanoQC (De Coster et al. 2018)), critical first steps in evaluating lrRNA-<br>seq read quality. Other tools, such as SQAN (Fukasawa et al. 2020), nanoQC (De Coster et al. 2018)), critical first steps in evaluating lrRNA-<br>seq read quality. Other tools, such as SQANTI3 (Pardo-Palacios et al. 2024a), TALON (Wyman et<br>al. 2020), FLAME (Holmqvist e most current tools for lrRNA-seq read quality control were developed during the early stages of seq read quality. Other tools, such as SQANTI3 (Param Paramer Paramer Paramer Computer (Paramer 2014)<br>al. 2020), FLAME (Holmqvist et al. 2021), IsoSeq (https://isoseq.how/), and IsoTools (Lienhard<br>et al. 2023), focus on ev al. 2023), focus on evaluating consensus transcript models inferred from the data. However,<br>most current tools for IrRNA-seq read quality control were developed during the early stages of<br>these technologies and are general et al. 2023), some calculations of the allegations were developed during the early stages of<br>these technologies and are generally limited in the number of evaluated features and/or<br>samples. As long-read sequencing technolo these technologies and are generally limited in the number of evaluated features and/or<br>samples. As long-read sequencing technologies rapidly evolve, improving both in quality and<br>experimental scope, the need for comprehen samples. As long-read sequencing technologies rapidly evolve, improving both in quality and experimental scope, the need for comprehensive and comparative read quality assessment<br>becomes increasingly critical.<br>The rapid decline in costs implies that the use of IrRNA-seq will continue to expand, with

experimental scope, the need for comprehensive and comparative read quality assessment<br>becomes increasingly critical.<br>The rapid decline in costs implies that the use of IrRNA-seq will continue to expand, with<br>experimental becomes in the rapid decline in costs in<br>experimental designs involvir  $\begin{array}{c}\n1 \\
1\n\end{array}$ experimental designs involving multiple samples becoming more common (e.g. (Glinos et al. experimental designs involving multiple samples becoming more common (e.g. (Glinos et al.

2022; Joglekar et al. 2024; Mahmoud et al. 2024; Patowary et al. 2024)). From a quality control<br>perspective, this necessitates that datasets be homogeneous, without biases associated with<br>experimental groups, and free of o experimental groups, and free of outliers. Moreover, the generated data must be sufficient to<br>address the research questions that motivated the experiment. The increase in throughput now<br>makes it possible to design experim address the research questions that motivated the experiment. The increase in throughput now<br>makes it possible to design experiments that include barcoding and multiplexing to balance<br>library preparation and sequencing acr makes it possible to design experiments that include barcoding and multiplexing to balance<br>library preparation and sequencing across experimental groups (Auer and Doerge 2010). This<br>approach helps avoid confounding technic ibrary preparation and sequencing across experimental groups (Auer and Doerge 2010). This<br>approach helps avoid confounding technical variation with the treatments of interest and<br>facilitates the identification of failed te approach helps avoid confounding technical variation with the treatments of interest and<br>facilitates the identification of failed technical replicates versus failed samples. Finally,<br>technological advancements such as more facilitates the identification of failed technical replicates versus failed samples. Finally,<br>technological advancements such as more accurate basecallers<br>(https://github.com/nanoporetech/dorado) and novel library preparat Fachnological advancements such as more accurate basecallers<br>
<u>(https://github.com/nanoporetech/dorado</u>) and novel library preparation methods (e.g. MAS-<br>
Iso-Seq (Al'Khafaji et al. 2024), CapTrap(Carbonell-Sala et al. 202 om (https://github.com/nanoporetech/dorado) and novel library preparation methods (e.g. MAS-<br>Iso-Seq (Al'Khafaji et al. 2024), CapTrap(Carbonell-Sala et al. 2024), and R2C2(Volden et al.<br>2018), Nano3P-seq (Begik et al. 202 Iso-Seq (Altion 1999), Nano3P-seq (Begik et al. 2023) FLAM-seq (Legnini et al. 2019)) require tools that can<br>
Leasily evaluate how these improvements impact various aspects of data quality.<br>
In this context, we present SQA

2024), Nanos Pesq (Begin et al. 2027, PERM-sequence al. 2024), Pequine Centeralistical<br>2024)<br>2024 Flame Scheel (Begin et al. 2024)<br>2024 Flame Scheel (Beginni et al. 2024), a tool originally designed for transcript model qu easily evaluation of the this context, we present SQANTI-reads, an extension of SQANTI3 (Pardo<br>2024a), a tool originally designed for transcript model quality control, to jointly<br>control metrics for long read data and to a  $\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array}$ 2024a), a tool originally designed for transcript model quality control, to jointly provide quality<br>control metrics for long read data and to analyze multiple samples for consistency and bias. We<br>demonstrate that SQANTI3's 2024 a)<br>2021 - Control metrics for long read data and to analyze multiple samples for consistency and bias. We<br>2021 demonstrate that SQANTI3's structural categories and other quality control metrics, repurposed<br>2021 in SQA control metrics for long read data and to analyze multiple samples for lementary and data is a demonstrate that SQANTI3's structural categories and other quality control metrics, repurposed<br>in SQANTI-reads, are highly effe demonstrate that SQANTI3's structural categories and other quality control metrics, repurposed<br>in SQANTI-reads, are highly effective for assessing the consistency of a IrRNA-seq multi-sample<br>experiment, identifying read qu experiment, identifying read quality control failures, and detecting outliers. Additionally, we<br>have developed new metrics that provide insights into the potential utility and discovery power<br>of the data, including variati have developed new metrics that provide insights into the potential utility and discovery power<br>of the data, including variation at donor/acceptor sites and identification of potentially under-<br>annotated genes and mis-anno of the data, including variation at donor/acceptor sites and identification of potentially under-<br>annotated genes and mis-annotated transcripts. SQANTI-reads offers an extensive array of<br>summary output tables, is customiza annotated genes and mis-annotated transcripts. SQANTI-reads offer an enterties analy stranscripts.<br>summary output tables, is customizable to accommodate any experimental design, and is<br>available as an open-source, freely a summary of the customizable tables is customized.<br>
Someon of the same open-source, freely accessible tool.<br>
The common open-source, freely accessible tool.<br>
The common of the common of the common of the common of the commo available as an open-source, freely accessible tool.

#### SQANTI-reads basics

METHODS<br>SQANTI-reads basics<br>SQANTI-reads is an adaptation of SQANTI3 designed to evaluate individual reads rather than<br>transcript models. It allows for the comparison of multiple samples, providing quality control<br>results transcript models. It allows for the comparison of multiple samples, providing quality control<br>results across the entire experiment. Several new features have been introduced to address the<br>specific needs of QC in multi-sa The comparts across the entire experiment. Several new features have been introduced to address the<br>specific needs of QC in multi-sample experiments, while some functionalities of SQANTI3 have<br>been removed as they are not result across the entire experiments across the entire experiments of SQANTI3 have<br>been removed as they are not applicable to read-level processing. Table 1 highlights the major<br>differences between SQANTI3 and SQANTI-reads specific needs of the variance of applicable to read-level processing. Table 1 highlights the major<br>differences between SQANTI3 and SQANTI-reads, emphasizing the new features of SQANTI-<br>reads, while Table 2 lists the names differences between SQANTI3 and SQANTI-reads, emphasizing the new features of SQANTI-reads, while Table 2 lists the names and descriptions of the output files.

<b>Feature</b>	<b>SQANTI3</b>	<b>SQANTI-reads</b>
Sequences analyzed	transcript models	reads
Annotation with SQANTI3 categories	yes, for transcript models	yes, for reads and UJCs
Computation of SQANTI3 quality metrics	yes	yes
Samples processed	One	Multiple
Visualizations across samples	no	yes
PCA analysis between samples	no	yes
Summary of read counts (per gene, per UJC)	no	yes
Donor/acceptor variation metrics	no	yes
Identification of putative under annotated genes	no	yes
Identification of putative novel transcripts	no	yes
Machine learning validation of transcript models	yes	no
<b>IsoAnnot annotation</b>	yes	no

Table 1. Comparison between SOANTI3 and SOANTI-reads

Table 2. SQANTI-reads specific output files





 $\frac{1}{2}$ genome FASTA file, 3) a GTF file of the reference transcript model annotation, and 4) a design<br>file containing metadata for multiple samples. The first step of SQANTI-reads involves using the<br>SQANTI3-QC module to generate file containing metadata for multiple samples. The first step of SQANTI-reads involves using the<br>SQANTI3-QC module to generate SQANTI3-like classification and junction files, with the<br>classification file containing one row SQANTI3-QC module to generate SQANTI3-like classification and junction files, with the<br>classification file containing one row for each mapped read. Reads are classified according to<br>the SQANTI categories (Tardaguila et al. Classification file containing one row for each mapped read. Reads are classified according to<br>the SQANTI categories (Tardaguila et al. 2018) as full-splice match (FSM), incomplete-splice<br>match (ISM), novel-in-catalog (NIC classification file containing the containing process in the contain mapped read. The SQANTI categories (Tardaguila et al. 2018) as full-splice match (FSM), incomplete-splice match (ISM), novel-in-catalog (NIC), novel-notmatch (ISM), novel-in-catalog (NIC), novel-not-in-catalog (NNC), antisense, fusion, genic<br>genomic, and intergenic. SQANTI3 subcategories are also included, based on 5' and 3' end<br>positions relative to the annotated transcr match (CM), increasing (NIC), increasing the catalog (NIC), and cateally carry, generical<br>genomic, and intergenic. SQANTI3 subcategories are also included, based on 5' and 3' end<br>positions relative to the annotated transcr genomic, and intergenic. SQANTI3 subcategories are also included, based on 5' and 3' end<br>positions relative to the annotated transcription start sites (TSS) and transcription termination<br>sites (TTS) (Pardo-Palacios et al. prelative to the annotated transcription start sites (TTS) (Pardo-Palacios et al. 2024a). Additionally, the reverse transcriptase (RT) switching algorithm of SQANTI3-QC identifies reads with evidence of RT switching events sites (TRS) algorithm of SQANTI3-QC identifies reads with evidence of RT switching events, while reads<br>with more than 60% adenines in the 20 bp downstream of the reported TTS at the genomic<br>level are flagged as potential i with more than 60% adenines in the 20 bp downstream of the reported TTS at the genomic<br>level are flagged as potential intrapriming events. The length of each read and the number of<br>exons in each read are also recorded in t level are flagged as potential intrapriming events. The length of each read and the number of<br>exons in each read are also recorded in the classification file.<br>Junction Metrics lexons in each read are also recorded in the classification file.<br> **Junction Metrics<br>
The SQANTI-reads junction file follows the same format as the SQANTI3 junction file, with each**  $2$ ANTI-reads

exons in each read are also recorded in the classification with<br>Junction Metrics<br>The SQANTI-reads junction file follows the same format as th<br>row representing a junction in a read, including the start and Junction Metrics<br>The SQANTI-read<br>row representing row representing a junction in a read, including the start and end positions of the junction. The row representing a junction in a read, including the start and end positions of the junction. The

distance from the junction start and end to the nearest annotated junction start and end in the reference GTF is calculated. It's important to note that the nearest annotated start and end positions may not belong to the s reference GTF is calculated. It's important to note that the nearest annotated start and end<br>positions may not belong to the same annotated junction. SQANTI classifies junctions as known<br>or novel, and as canonical or non-c por novel, and as canonical or non-canonical, based on the dinucleotide pairs at the junction's<br>start and end. By default, dinucleotide combinations of GT-AG, GC-AG, and AT-AC are<br>considered canonical, while any other comb or novel, and as canonical or non-canonical, based on the dinucleotide pairs at the junction's<br>start and end. By default, dinucleotide combinations of GT-AG, GC-AG, and AT-AC are<br>considered canonical, while any considered canonical, while any other combinations are classified as non-canonical, although<br>the user can specify additional canonical sites.

consider user can specify additional canonical sites.<br>SQANTI-reads introduces new metrics to evaluate the relationship between the junctions in<br>The square combination as and the annotated donors and acceptors. In the SQANT the user can specify additional canonical cases.<br>SQANTI-reads introduces new metrics to eva<br>mapped reads and the annotated donors and<br>distance from each donor/acceptor in each re S<br>r<br>r mapped reads and the annotated donors and acceptors. In the SQANTI3-QC junction file, the<br>distance from each donor/acceptor in each read to the nearest annotated donor/acceptor is<br>recorded. In SQANTI-reads, the mean absolu mapped reads and an amountation and acceptors in the SQANTI  $\alpha$  junction  $\alpha$ ) in distance from each donor/acceptor is<br>recorded. In SQANTI-reads, the mean absolute distance in nucleotides from the annotated<br>donor/acceptor recorded. In SQANTI-reads, the mean absolute distance in nucleotides from the annotated<br>donor/acceptor site, the standard deviation, and the coefficient of variation (CV = standard<br>deviation/mean) are calculated and includ donor/acceptor site, the standard deviation, and the coefficient of variation (CV = standard deviation/mean) are calculated and included in the cv.csv file. Each detected junction is classified as 1) Reference Match junction if the mean distance and the standard deviation to an annotated junction are both equal to 0; 2) CV = 0 Junction when the mean distance is greater than 0 and the standard deviatio annotated junction are both equal to 0; 2) CV = 0 Junction when the mean distance is greater than 0 and the standard deviation equals 0, and 3) CV  $>$  0 Junction when the CV is greater than

## Unique Junction Chain and gene-level information

SQANTI-reads groups mapped reads based on their junctions, referred to as Unique Junction Ur<br>SC<br>Ch **しょうしょう** Chains (UJCs). Each UJC is labeled with a string that includes the chromosome and junction coordinates (Nanni et al. 2024). To enhance computational efficiency, UJC strings are encoded as an index in a hash table (JxnHash) coordinates (Nanni et al. 2024). To enhance computational efficiency, UJC strings are encoded<br>as an index in a hash table (JxnHash). The read count for each JxnHash is calculated and<br>included in the ujc\_counts.csv file. Ad as an index in a hash table (JxnHash). The read count for each JxnHash is calculated and<br>included in the ujc\_counts.csv file. Additionally, the number of known canonical, known non-<br>canonical, novel canonical, and novel no included in the ujc\_counts.csv file. Additionally, the number of known canonical, known non-<br>canonical, novel canonical, and novel non-canonical junctions within each UJC is annotated,<br>along with the SQANTI structural cate summary file gene counts.csv. along with the SQANTI structural category of the UJC. The number of reads within each<br>structural category for each gene, as well as the total number of reads per gene, is stored in the<br>summary file gene\_counts.csv. structural category for each gene, as well as the total number of reads per gene, is stored in the summary file gene<sub>,</sub> counts.csv. summary file generalized and counts.csv.

#### Identifying genes that may be under-annotated and transcripts that may be mis-annotated

|
|
|
| The presence of a potentially novel transcript. SQANTI-reads includes a customizable pipeline to<br>identify genes with such potential under-annotation events. The procedure identifies NIC/NNC<br>UJCs that meet a minimum number the presence of a potential under-annotation events. The procedure identifies NIC/NNC<br>UJCs that meet a minimum number (R) and proportion (P) of reads, with default values set at<br>100 reads and 20%, respectively. To mitigate IDES that meet a minimum number (R) and proportion (P) of reads, with default values set at<br>100 reads and 20%, respectively. To mitigate the risk that the NIC/NNC UJC is merely a<br>degradation product, an additional conditio 100 reads and 20%, respectively. To mitigate the risk that the NIC/NNC UJC is merely a<br>degradation product, an additional condition is applied: the candidate UJC must include at least<br>80% of the gene's junctions (Q) (Figur degradation product, an additional condition is applied: the candidate UJC must include at least<br>80% of the gene's junctions (Q) (Figure 3A). The R, P, and Q thresholds are pipeline parameters<br>that can be adjusted by the u  $\frac{1}{2}$ <br>80% of the gene's junctions (Q) (Figure 3A). The R, P, and Q thresholds are pipeline parameters<br>that can be adjusted by the user. Furthermore, SQANTI-reads allows for the evaluation of<br>under-annotated genes and 80% of the gene's junctions (Q) (Figure 3A). The R, P, and Q thresholds are pipeline parameters<br>that can be adjusted by the user. Furthermore, SQANTI-reads allows for the evaluation of<br>under-annotated genes and novel trans that can be adjacent by the user. Furthermore, SQANTI-reads antite the unit of the evaluation of<br>under-annotated genes and novel transcripts within a specific subset of samples associated<br>with a particular experimental fac with a particular experimental factor (e.g., developmental stage or technology) using the --<br>factor-level option.<br>Multisample processing factor-level option.<br>Wultisample processing<br>SQANTI-reads processes multiple samples to generate classification and junction files when a

## Multisample processing

ractor-lever option.<br>Multisample proce:<br>SQANTI-reads proce<br>design file (Suppler design file (Supplementary file 1) is provided to the sqanti-reads.py command. If individual<br>samples have already been pre-processed with SQANTI3-QC, SQANTI-reads can be run in --fast<br>mode, where the design file links the samples have already been pre-processed with SQANTI3-QC, SQANTI-reads can be run in --fast samples have already mode, where the design file links the individual classification and junction files to sample IDs for<br>the calculation of SQANTI-reads metrics, summaries, and a series of visualizations. If pre-<br>processi mode, where the design file links the distribution file individual classifications. If pre-<br>processing has not been done, SQANTI-reads is run in --simple mode where SQANTI3 is run on<br>each sample, followed by the calculatio quartile of mapped read length, as well as the number and proportion of reads that are shorter processing the increased by the calculation of SQANTI3 metrics and summaries. The output also<br>includes a summary for each sample, reporting the mean, median, upper quartile, and lower<br>quartile of mapped read length, as wel includes a summary for each sample, reporting the mean, median, upper quartile, and lower<br>quartile of mapped read length, as well as the number and proportion of reads that are shorter<br>than 1 kb, between 1 and 2 kb, betwee quartile of mapped read length, as well as the number and proportion of reads that are shorter<br>than 1 kb, between 1 and 2 kb, between 2 and 3 kb, and greater than 3 kb in length, all of which<br>are included in the length\_sum than 1 kb, between 1 and 2 kb, between 2 and 3 kb, and greater than 3 kb in length, all of which<br>are included in the length summary.csv file. than 2 kb, between 2 kb, between 2 kb, and 3 kb, greater than 2 kb in length, and which<br>are included in the length\_summary.csv file.<br>Drosophila melanogaster data<br>A total of 24 samples corresponding to 2 developmental stage

## $\overline{a}$ Drosophila melanogaster data

are all on-<br>Drosophila melanogaster data<br>A total of 24 samples corresponding to 2 de<br>hatching) and four genotypes (dmel 110 hatching) and four genotypes (dmel 11037, 11255, 12272 and 12279) (3 samples per experimental condition) were sequenced using Oxford Nanopore Technology (ONT). One<br>barcoded cDNA library was built per sample, the 12 samples from the 0-1 hour and the 12<br>samples from the 3-8 day stages were pooled and seq samples from the 3-8 day stages were pooled and sequenced on a MinION. Data were evaluated with PycoQC (Leger and Leonardi 2019) which focuses on read length and base<br>quality. All samples passed this basic QC, libraries were re-pooled and run on the PromethION.<br>Detailed metadata for these samples are p quality. All samples passed this basic QC, libraries were re-pooled and run on the PromethION.<br>Detailed metadata for these samples are provided in Supplementary Table 1. Raw electrical<br>data were processed by default baseca Detailed metadata for these samples are provided in Supplementary Table 1. Raw electrical<br>data were processed by default basecallers on the machines during the run (Guppy). Samples<br>were stored in .fast5 format and as .fast data were processed by default basecallers on the machines during the run (Guppy). Samples<br>were stored in .fast5 format and as .fastq files. The .fast5 files were converted to the Dorado<br>compatible .pod5 format using pod5 were stored in .fast5 format and as .fastq files. The .fast5 files were converted to the Dorado<br>compatible .pod5 format using pod5 (v 0.3.6) and then processed in Dorado (v 0.5.2)<br>(https://github.com/nanoporetech/dorado) u compatible .pod5 format using pod5 (v 0.3.6) and then processed in Dorado (v 0.5.2)<br>(https://github.com/nanoporetech/dorado) using options --recursive --device "cuda:0,1" --kit-<br>name SQK-PCB109 --trim none. Reads were demu 0.5.2) with options --no-classify --emit-fastq resulting in separate Dorado fastq files for each<br>sample. The fastq files generated by Guppy/Dorado were both processed using pychopper (v (heta ordentianoporthere in the sweep aligned to D. melanogaster 6.50 and the resulting sam files name SCA-2) with options --no-classify --emit-fastq resulting in separate Dorado fastq files for each<br>sample. The fastq files generated by Guppy/Dorado were both processed using pychopper (v<br>2.7.1), the oriented fastq fil sample. The fastq files generated by Guppy/Dorado were both processed using pychopper (v<br>2.7.1), the oriented fastq files were aligned to *D. melanogaster* 6.50 and the resulting sam files<br>were converted to gtf using samto 2.7.1), the oriented fastq files were aligned to *D. melanogaster* 6.50 and the resulting sam files<br>were converted to gtf using samtools (v 1.10) (Li et al. 2009) and bedtools (v 2.29.2) (Quinlan<br>and Hall 2010). The resul 2.7.1), the oriented fastq files were aligned to D. melanogaster 6.50 and the resulting sam files<br>were converted to gtf using samtools (v 1.10) (Li et al. 2009) and bedtools (v 2.29.2) (Quinlan<br>and Hall 2010). The resultin and Hall 2010). The resulting gtf files (67 technical replicates from 24 samples), the *D.*<br>
melanogaster 6.50 fasta reference file<br>
(https://ftp.flybase.net/releases/FB2023 01/dmel r6.50/ (Öztürk-Çolak et al. 2024)), and and Hall 2010). The resulting gtraines (or technical replicates from 24 samples), the *D.*<br>
melanogaster 6.50 fasta reference file<br>
(https://ftp.flybase.net/releases/FB2023 01/dmel r6.50/ (Öztürk-Çolak et al. 2024)), and a melanogaster b.50 asstalled that the reference<br>(https://ftp.flybase.net/releases/FB2023 01/dmel r6.50/ (Öztürk-Çolak et al. 2024)), and a<br>design file (Supplementary File 2) were used as input to SQANTI-reads. The *Drosophi* (design file (Supplementary File 2) were used as input to SQANTI-reads. The *Drosophila* dataset<br>includes, therefore, two experimental conditions (time and genotype) and two technical<br>conditions (sequencing platform and ba design the (Supplementary File 2) were used as input to SQANTI-reads. The *Drosophila* dataset<br>includes, therefore, two experimental conditions (time and genotype) and two technical<br>conditions (sequencing platform and base conditions (sequencing platform and base caller) with the experimental samples multiplexed<br>and evaluated with both technical conditions. The SQANTI-reads output for this dataset is<br>provided in Supplementary File 3. and evaluated with both technical conditions. The SQANTI-reads output for this dataset is provided in Supplementary File 3. provided in Supplementary File 3.<br>And the SQANTI-read of the SQANTI-read RNA-seq<br>The SQANTI-read RNA-seq output for the SQANTI-read RNA-seq Genome Annotation<br>We used publicly available lrRNA-seq data from the Long-read RNA

## Human Cell line WTC11

provided in Supprovidentary File 3.<br>Human Cell line WTC11<br>We used publicly available lrRNA<br>Assessment Project (Pardo-Palac ーーノ Assessment Project (Pardo-Palacios et al. 2024b) to illustrate the utility of SQANTI-reads.<br>Specifically, we used triplicate measurements of the transcriptome of the WTC11 human cell<br>line that were profiled by cDNA PacBio Specifically, we used triplicate measurements of the transcriptome of the WTC11 human cell<br>line that were profiled by cDNA PacBio Sequel II, cDNA Oxford Nanopore Minion, and direct line that were profiled by cDNA PacBio Sequel II, cDNA Oxford Nanopore Minion, and direct line that were profiled by cDNA PacBio Sequel II, cDNA Oxford Nanopore Minion, and direct

Website (https://www.encodeproject.org/search/?type=Experiment&internal tags=LRGASP).<br>Accession numbers for these samples are provided in Supplementary Table 2. The fastq files<br>were pre-processed by LRGASP researchers as d Accession numbers for these samples are provided in Supplementary Table 2. The fastq files<br>were pre-processed by LRGASP researchers as described in (Pardo-Palacios et al. 2024b). We<br>used the gtf files of read alignments, G were pre-processed by LRGASP researchers as described in (Pardo-Palacios et al. 2024b). We<br>used the gtf files of read alignments, GENCODE's GRCh38.p13 reference genome gtf and fasta<br>for release 38 (https://www.gencodegenes used the gtf files of read alignments, GENCODE's GRCh38.p13 reference genome gtf and fasta<br>for release 38 (https://www.gencodegenes.org/human/release 38.html ), and a design file<br>(Supplementary file 4) to run SQANTI-reads for release 38 (https://www.gencodegenes.org/human/release 38.html ), and a design file<br>(Supplementary file 4) to run SQANTI-reads on the WTC11 samples. The SQANTI-reads output is<br>provided in Supplementary File 5. (Supplementary file 4) to run SQANTI-reads on the WTC11 samples. The SQANTI-reads output is<br>provided in Supplementary File 5. (Supplementary file 4) to run SQANTI-reads on the WTCL1 samples on the WTC11 samples.<br>provided in Supplementary File 5.<br>RESULTS provided in Supplementary File 5.<br>RESULTS<br>SQANTI-reads can be used to evaluate technology for a common set of libraries

#### RESULTS

 ו<br>ג<br>ה )<br>1<br>1 SQANTI-reads can be used to evaluate technology for a common set of instants<br>Our *Drosophila* experimental design serves as an excellent example to<br>technological aspects of long-read methods using SQANTI-reads. One of<br>adva Our *Brosophila* experimental design serves as an excellent example to evaluate the<br>technological aspects of long-read methods using SQANTI-reads. One of the notable<br>advancements in Oxford Nanopore Technology (ONT) is the advancements in Oxford Nanopore Technology (ONT) is the introduction of the new basecaller,<br>Dorado (https://github.com/nanoporetech/dorado), which claimed to significantly improve<br>base call accuracy. A robust long-read QC Dorado (https://github.com/nanoporetech/dorado), which claimed to significantly improve<br>base call accuracy. A robust long-read QC tool should be able to clearly identify improvements<br>in technology. Therefore, we first comp base call accuracy. A robust long-read QC tool should be able to clearly identify improvements<br>in technology. Therefore, we first compared the Guppy and Dorado basecallers using SQANTI-<br>reads. As anticipated, Dorado result in technology. Therefore, we first compared the Guppy and Dorado basecallers using SQANTI-<br>reads. As anticipated, Dorado resulted in more reads with assignable barcodes, a higher<br>number of mapped reads, more reads aligning (Supplementary Figure 1, Supplementary file 2). This confirms that Dorado improves basereads of mapped reads, more reads aligning to annotated genes, and more reads aligning to<br>annotated transcripts, without an increase in the proportion of reads with technical artifacts<br>(Supplementary Figure 1, Supplementar annotated transcripts, without an increase in the proportion of reads with technical artifacts<br>(Supplementary Figure 1, Supplementary file 2). This confirms that Dorado improves base-<br>calling accuracy without introducing u calling accuracy without introducing unwanted biases. Based on these SQANTI-reads QC results<br>we chose to move forward with Dorado basecalled reads.

In the Drosophila experiment, libraries were barcoded, pooled, and multiplexed across different calling accuracy introducts and statement and these statements and these SQANTI-reads with the chose to move forward with Dorado basecalled reads.<br>In the *Drosophila* experiment, libraries were barcoded, pooled, and multip In the *Drosophila* experiment, libraries were barcoded, pooled,<br>MinION and PromethION runs, with a re-pooling step betwe<br>SQANTI-reads to compare the quality of the MinION and the P MinION and PromethION runs, with a re-pooling step between the two machines. We used SQANTI-reads to compare the quality of the MinION and the PromethION runs, and to evaluate the consistency of the PromethION technology across technical replicates. The first MinION run the construction of the Promethion technology across technology across technical replications. The first MinIO

(TR1) had higher percentages of reads with NIC/NNC and with non-canonical junctions<br>compared to the second and third technical MinION runs (TR2,TR3) and compared to the<br>PromethION runs for the same libraries (Supplementary PromethION runs for the same libraries (Supplementary Figure 2). The other MinION runs (TR2,<br>TR3) were similar in their quality metrics (described below) to the PromethION run of the same<br>samples, and the technical replica TR3) were similar in their quality metrics (described below) to the PromethION run of the same<br>samples, and the technical replicates of the 3-8 day libraries on the PromethION were similar.<br>These results indicate that the TR3) were similar in their quality metric (described below) to the PromethION were similar.<br>These results indicate that the technology performs consistently across instruments and runs.<br>Based on these SQANTI-reads QC resul These results indicate that the technology performs consistently across instruments and runs.<br>Based on these SQANTI-reads QC results we aggregated data across technical replicates to<br>further evaluate the quality of the IrR Based on these SQANTI-reads QC results we aggregated data across technical replicates to<br>further evaluate the quality of the lrRNA-seq experiment.

In a multi-sample IrRNAseg experiment, all samples should be of similar quality. SQANTI3 uses further and the lines.<br>SQANTI-reads metrics can be used to evaluate the global<br>In a multi-sample lrRNAseq experiment, all samples shou<br>the FSM structural category to identify long-read seque S<br>I<br>I<br>I SQANTI-reads metrics can be used to evaluate the global quality of the lrRNA-seq experiment In a FSM structural category to identify long-read sequences whose junctions are consistent<br>In an annotated transcript model. However, for a IrRNA-seq experiment to accurately reflect<br>In analyzed transcriptome, the reads s with an annotated transcript model. However, for a lrRNA-seq experiment to accurately reflect the analyzed transcriptome, the reads should ideally also capture the distribution of transcript lengths of the expressed transcriptome. The distribution of transcript lengths depends on the species with *Drosophila* having overall less complex and shorter transcripts than human<br>(Supplementary Figure 3). A dataset with reads substantially shorter than the targeted<br>transcriptome but with still a high number of (Supplementary Figure 3). A dataset with reads substantially shorter than the targeted transcriptome but with still a high number of FSM indicates capture of short transcripts, while combining shorter than expected reads with a high proportion of ISM may indicate RNA degradation. We looked at these values fo transcription of the combining shorter than expected reads with a high proportion of ISM may indicate RNA<br>degradation. We looked at these values for an initial assessment of the quality of the<br>Drosophila experiment. combining shorter than expected reads with a high proportion of the quality of the<br>degradation. We looked at these values for an initial assessment of the quality of the<br>Drosophila experiment.<br>First, we compared the number

degroophila experiment.<br>Drosophila experiment.<br>First, we compared the number of reads and length distributions for all samples. The difference<br>in sequencing depth between the two developmental stages was evident (Figure 1A Brosophila experiment.<br>First, we compared the<br>in sequencing depth be<br>samples, most reads w li<br>i<br>t First, we compared the number of reads and length distribution and length of requencing depth between the two developmental stages was evident (Figure 1A). For all samples, most reads were shorter than 1kb with less than 2 samples, most reads were shorter than 1kb with less than 20% of them above the 1 kb<br>threshold (Figure 1B, Supplementary Figure 4A). While between 53% and 67% of the reads<br>across samples were classified as FSM, 20% to 38% w samples, most reads were chastified as FSM, 20% to 38% were labeled as ISM, and NIC/NNC were<br>across samples were classified as FSM, 20% to 38% were labeled as ISM, and NIC/NNC were<br>under 10% of the reads (Figure 1C). For t across samples were classified as FSM, 20% to 38% were labeled as ISM, and NIC/NNC were<br>under 10% of the reads (Figure 1C). For the reads greater than 1 kb, between 73% and 82% of<br>the reads were FSM while 10% and 18% were under 10% of the reads (Figure 1C). For the reads greater than 1 kb, between 73% and 82% of<br>the reads were FSM while 10% and 18% were ISM (Figure 1D). the reads were FSM while 10% and 18% were ISM (Figure 1D).<br>
For the reads were FSM while 10% and 18% were ISM (Figure 1D).  $\mathbf{r}^{\mathcal{L}}$ 

 $\begin{array}{c} \mathsf{T} \\ \mathsf{S} \\ \mathsf{S} \end{array}$ To further understand how read quality affects gene and transcript quantification, we examined these metrics aggregated by gene and UJC. We found that, despite the large sequencing depth differences between developmental s sequencing depth differences between developmental stages, the number of detected genes<br>was only slightly lower in the 0-1 h samples (Figure 1E). However, these genes were quantified<br>with fewer reads (80% genes with < 50 r sequencing depth differences between developmental stages) are minimered as tensoring gener<br>was only slightly lower in the 0-1 h samples (Figure 1E). However, these genes were quantified<br>with fewer reads (80% genes with < with fewer reads (80% genes with < 50 reads) than the 38 d samples, which had between 30%<br>and 50% of genes with more than 100 reads(Figure 1E). Interestingly, when evaluating UJC we<br>found that, while the number of UJC mirr and 50% of genes with more than 100 reads(Figure 1E). Interestingly, when evaluating UJC we<br>found that, while the number of UJC mirrored the sequencing depth pattern (Figure 1F), with 3-<br>8 d samples showing five times more found that, while the number of UJC mirrored the sequencing depth pattern (Figure 1F), with 3-<br>8 d samples showing five times more UJC than 1 h samples, and a larger number of FSM and<br>ISM UJC, there were many additional UJ Found is a samples showing five times more UJC than 1 h samples, and a larger number of FSM and<br>ISM UJC, there were many additional UJC detected by fewer than 10 reads, and usually by a<br>Single read (Figure 1F & 1G) and the ISM UJC, there were many additional UJC detected by fewer than 10 reads, and usually by a<br>single read (Figure 1F & 1G) and these UJC were most frequently NIC/NNC (Supplementary<br>Figure 4B & 4C). Downstream analyses would th ISM 1997 MELL MELL MANY ADDITIONAL PREPERTENT AND THE MUMICLE CERTY AND MELL MY AND ANY OF Single read (Figure 4B & 4C). Downstream analyses would therefore need to address whether this represents novel low-expressed trans Figure 4B & 4C). Downstream analyses would therefore need to address whether this<br>represents novel low-expressed transcripts or technology errors. In contrast, the percentage of<br>FSM reads between the two time points differ Figure 4B & 4C). Determined analyses to them analyses where the additional interest and<br>represents novel low-expressed transcripts or technology errors. In contrast, the percentage of<br>FSM reads between the two time points FSM reads between the two time points differed by less than 1x in all replicates (Figure 1H).<br>These results indicate that the higher sequencing depth of the 3-8 d samples does not change<br>the number of detected genes or ann These results indicate that the higher sequencing depth of the 3-8 d samples does not change<br>the number of detected genes or annotated transcripts (FSM). The higher read depth per gene/<br>UJC suggests that more genes and tra The number of detected genes or annotated transcripts (FSM). The higher read depth per gene/<br>UJC suggests that more genes and transcripts will be able to be quantitatively evaluated in the<br>3-8 day samples compared to the 0 UJC suggests that more genes and transcripts will be able to be quantitatively evaluated in the 3-8 day samples compared to the 0-1 hour samples.

In the *Drosophila* data, we noticed two samples (RIL 12279 rep 1 0-1 hour- orange arrow; RIL<br>11255 3-8 day rep 3 - teal arrow) that had the lowest percentage of FSM and highest percentage of ISM in the 0-1 hour and 3-8 day groups respectively (Figure 1C). To determine |
|
|
|
| In the *Brosophila* data, we noticed two samples (RIL 12279 rep 1 0-1 hour- orange arrow, RIL<br>11255–3-8 day rep 3 - teal arrow) that had the lowest percentage of FSM and highest<br>percentage of ISM in the 0-1 hour and 3-8 da 1125 1125 day rep 1125 and the lower and 3-8 day groups respectively (Figure 1C). To determine<br>whether these two samples were of overall lower quality than the rest, we examined the<br>SQANTI-reads metrics for these two sampl percentage of Islam in the 1-1 hours and 1-1 hours groups respectively (Figure 10). To determine<br>whether these two samples were of overall lower quality than the rest, we examined the<br>SQANTI-reads metrics for these two sam SQANTI-reads metrics for these two samples. We found that RIL 12279 rep 1 had a lower<br>proportion of FSM across all genes (Figure 1H) and a higher proportion of genes quantified with<br>only one gene (Figure 1F), while RIL 112 proportion of FSM across all genes (Figure 1H) and a higher proportion of genes quantified with<br>only one gene (Figure 1F), while RIL 11255 rep 3 had a similar gene (Figure 1H), UJC (Figure 1G)<br>and % FSM in genes (Figure 1H only one gene (Figure 1F), while RIL 11255 rep 3 had a similar gene (Figure 1H), UJC (Figure 1G) and % FSM in genes (Figure 1H) than other 3-8 day samples. We concluded that RIL 12279 rep 1 and  $W$  FSM in gener (Figure 1H) than other 3-8 day samples. We concluded that RIL 12279  $P$  = 0-1 hour is a low-quality sample. 0-1 hour is a low-quality sample.

Altogether, this example shows that SQANTI-reads metrics can be used to compare samples s and experimental conditions in a multi-sample experiment, detect outliers, and suggest points s of attention for downstream data processing.



Figure 1: SQANTI-reads analysis of Drosophila samples. A) Number of mapped reads by experimental group Figure 1: SQANTI-reads analysis of Drosophila samples. A) Number of mapped reads by experimental group<br>Is below with meal langth. B) Bencenters of managemental groups in the managemental group of the CCM. But a management labeled with read length. B) Percentage of mapped reads >1kb vs percentage of reads that are FSM. Dots represent t early stage (0-1 hours after enclosure) and crosses indicate adult stage (3 to 8 days old). The four genotypes are indicated with four different colors. C) Percentage of reads mapping to genes in each SQANTI3 structural category y D) Percentage of reads mapping to genes in each SQANTI3-QC structural category for reads >1kb E) Number of the<br>Form of the structural structural structural category for reads and the structural category for reads +1kb E) N genes detected with breakdown by the number of reads mapped to each gene. F) Number of Ostal and the number of breakdown by the number of reads associated with each UJC. G) Proportion of USS detected with breakdown by the number of reads associated with each UJC. H) Distribution of the percentage of FSM reads by gene across samples. y<br>. s

#### SQANTI-reads metrics can be used to identify systematic differences among samples

The previous example according time of the previous are the total differences making it<br>dataset consistency. However, SQANTI-reads evaluates over 35 quality metrics, making it<br>challenging to determine which features contri data in the challenging to determine which features contribute to potential differences among samples.<br>We include Principal Component Analysis (PCA) analysis to identify which metrics are the most<br>relevant for quality vari challenging to determine minimized contribute to potential differences amorg samples are the most<br>relevant for quality variability when there are differences among samples or between groups.<br>The percentage of reads and UJC relevant for quality variability when there are differences among samples or between groups.<br>The percentage of reads and UJCs in each structural category, percentage of artifact reads (RT-<br>switching, non-canonical junction The percentage of reads and UJCs in each structural category, percentage of artifact reads (RT-<br>switching, non-canonical junctions and intrapriming), percentage of junctions in each category,<br>as well as length metrics, are

We applied SQANTI-reads PCA analysis of quality features to investigate differences in read swell as length metrics, are included in the PCA.<br>We applied SQANTI-reads PCA analysis of quality features to investigate differences in read<br>quality among various long-read sequencing methods used in the LRGASP challenge we applied SQANTI-reads PCA analysis of qualit<br>quality among various long-read sequencing me<br>Palacios et al. 2024b), focusing on the WTC11 \<br>C<br>S Palacios et al. 2024b), focusing on the WTC11 dataset. The analysis revealed that WTC11 samples clustered based on the long-read technology applied (Figure 2A). Specifically, PC1, which explains 56% of the variance, distin the other two technologies, while PC2, accounting for 35% of the variance, highlighted Sumples clustered based on the long-read technology applied (Figure 2A). Specifically, PC1,<br>Which explains 56% of the variance, distinguished cDNA ONT samples from those generated by<br>the other two technologies, while PC2, samples from those generated by<br>the other two technologies, while PC2, accounting for 35% of the variance, highlighted<br>differences between dRNA ONT and cDNA PacBio. To further explore these differences, we<br>examined the loa which explains 56% of the variance, highlighted the variance, distinguished differences between dRNA ONT and cDNA PacBio. To further explore these differences, we examined the loadings for each principal component. Quality differences between dRNA ONT and cDNA PacBio. To further explore these differences, we<br>examined the loadings for each principal component. Quality features with the highest positive<br>loadings in PC1 included the number of r examined the loadings for each principal component. Quality features with the highest positive<br>loadings in PC1 included the number of reads, the percentage of reads and the proportion of<br>UJCs in the NNC category, while fea examples and the proportion of loadings in PC1 included the number of reads, the percentage of reads and the proportion of<br>UJCs in the NNC category, while features with high negative loadings included Intergenic and<br>Genic LOCS in the NNC category, while features with high negative loadings included Intergenic and<br>Genic Genomic reads. Several junction-related variables also exhibited high absolute loadings<br>on PC1 (Figure 2B). SQANTI-reads pl UP TO CONDERVIES ON BOTH CONDERT CONDERTIGN CONTROLLER SURVEY OF STRING ON PC1 (Figure 2B). SQANTI-reads plots confirmed these structural category differences between cDNA ONT samples and other library preparations. cDNA O FENIT CHAING CENTRICITY JUNCTION-RELATION-RELATION-RELATION-RELATION-RELATION-RELATION-RELATION-RELATION-RELATION<br>Setween CDNA ONT samples and other library preparations. CDNA ONT had both the highest<br>proportion of NNC rea on PC1 (Figure 20). Such these precedinates and these structure between cDNA ONT samples and other library preparations. cDNA ONT had both the highest proportion of NNC reads and UJCs, (Figure 2C and 2D) and also had the l proportion of NNC reads and UJCs, (Figure 2C and 2D) and also had the lowest proportion of<br>intergenic reads (Figure 2C). Other differences in sequencing throughput and junction<br>characteristics were also confirmed (Suppleme intergenic reads (Figure 2C). Other differences in sequencing throughput and junction<br>characteristics were also confirmed (Supplementary Figure 5).

characteristics were also confirmed (Supplementary Figure 5).<br>Upon examining the feature loadings for PC2, we found that variables with high contributions<br>included several metrics related to read length (Figure 2B). Conseq Upon examining the feature loadings for PC2, we found that<br>included several metrics related to read length (Figure 2B).  $\frac{1}{i}$ included several metrics related to read length (Figure 2B). Consequently, we evaluated the included several metric relations relations relations relations  $\mathbb{R}^n$ . Consequently, we evaluated the

SQANTI-reads "Lengths of All Mapped Reads" plot for this experiment. Indeed, we observed d that the cDNA PacBio method produced a significantly higher proportion of reads between 1-2 2 kb, 2-3 kb, and greater than 3 kb, as suggested by their negative loadings, compared to the dDNA ONT method, which predominantly generated reads shorter than 1 kb (Figure 2E) . Similariy, percentage of reads assigned as ISM with high positive values was higher in driver ONT samples (Figure 2 C).

In conclusion, the SQANTI-reads PCA analysis is an effective tool for uncovering significant read  $\overline{a}$ quality differences between long-read sequencing methods and for revealing unexpected d technology biases.



 $\frac{1}{2}$  is  $\frac{1}{2}$  such the easy of LRGAS products  $\frac{1}{2}$  and  $\frac{1}{2}$  is  $\frac{1}{2}$  and  $\frac{1}{2}$  is  $\frac{1}{2}$  is  $\frac{1}{2}$  in  $\frac{1}{2}$  is  $\frac{1}{2}$  in  $\frac{1}{2}$  in  $\frac{1}{2}$  is  $\frac{1}{2}$  in  $\frac{1}{2}$  in  $\frac{1}{2}$ Top 10 Loadings for PC1 and PC2. C) Distribution of reads in SQANTI structural categories. D) Distribution of UJCs in

# structural categories E) Distribution of read lengths for all mapped reads.<br>SQANTI-reads identifies potentially under annotated genes.<br>Long-read data often reveal a large number of sequences t SQANTI-reads identifies potentially under annotated genes.

 $\frac{1}{2}$  $\frac{1}{1}$ S<br>I<br>∈ existing annotations. In many cases, these UJC belong to annotated genes and are identified by<br>only a few reads, as illustrated by the SQANTI-reads analysis in Figures 1 and 2. This suggests<br>they could be either low-expres exidency a few reads, as illustrated by the SQANTI-reads analysis in Figures 1 and 2. This suggests<br>they could be either low-expressed transcripts or technological artifacts. However, in some<br>instances, a high proportion o they could be either low-expressed transcripts or technological artifacts. However, in some<br>instances, a high proportion of reads in a gene may correspond to the same UJC, indicating the<br>possibility of a novel previously u instances, a high proportion of reads in a gene may correspond to the same UJC, indicating the<br>possibility of a novel previously unannotated transcript that warrants closer examination.<br>SQANTI-reads includes a customizable SQANTI-reads includes a customizable decision tree to identify such cases (see Methods, Figure<br>3A). We applied this approach to the WTC11 PacBio data and identified 8556 well annotated<br>genes, 88% of which have an annotated possiblity of a non-<br>SQANTI-reads includes a customizable decision tree to identify such cases (see Methods, Figure<br>3A). We applied this approach to the WTC11 PacBio data and identified 8556 well annotated<br>genes, 88% of wh SA). We applied this approach to the WTC11 PacBio data and identified 8556 well annotated<br>genes, 88% of which have an annotated transcript with high coverage (>20% of total gene<br>coverage) (Figure 3B). We also identified 10 states, 88% of which have an annotated transcript with high coverage (>20% of total gene<br>coverage) (Figure 3B). We also identified 101 underannotated genes, 54% of which have a<br>candidate unannotated transcript. The annotat genes, service in minimum and annotated manuscript minimigh coverage (service it man genes)<br>coverage) (Figure 3B). We also identified 101 underannotated genes, 54% of which have a<br>candidate unannotated transcript. The anno candidate unannotated transcript. The annotation category for all genes evaluated (default: R > 100) are provided in the gene\_classification.csv file. We also found 424 putative novel transcripts (Figure 3C). Of these, 316 candidation of the gene elassification csv file. We also found 424 putative novel<br>transcripts (Figure 3C). Of these, 316 were NIC and 108 were NNC transcripts. The SQANTI-<br>reads output for putative novel transcripts flagge transcripts (Figure 3C). Of these, 316 were NIC and 108 were NNC transcripts. The SQANTI-<br>reads output for putative novel transcripts flagged based on their length and read count are<br>output in the putative underannotation.

For genes with at least one putative novel transcript (R=100, P=20, Q=80) and an annotated reads surput for putative in the animary in 1952 marked on their length and read count are<br>output in the putative \_underannotation.csv file (Table 2).<br>For genes with at least one putative novel transcript (R=100, P=20, Q=8 out in the set of the putative novel transcript (R-<br>For genes with at least one putative novel transcript (R-<br>transcript (FSM) we selected the FSM with the highest pro<br>the structure of the putative novel transcripts to the ו<br>†<br>ד For genes with at least one putative in the highest proportion of reads. We then compared<br>the structure of the putative novel transcripts to the most expressed annotated transcript using<br>TranD (Nanni et al. 2024). For gene the structure of the putative novel transcripts to the most expressed annotated transcript using<br>TranD (Nanni et al. 2024). For genes with an annotated transcript that is relatively highly<br>expressed (>20% of the reads for TranD (Nanni et al. 2024). For genes with an annotated transcript that is relatively highly<br>expressed (>20% of the reads for that gene), 103 putative candidate transcripts differed from<br>the annotated transcript by donor/ac Expressed (>20% of the reads for that gene), 103 putative candidate transcripts differed from<br>the annotated transcript by donor/acceptor variation, suggesting a possible alternative splice<br>site. In addition, 10 putative ca the annotated transcript by donor/acceptor variation, suggesting a possible alternative splice<br>site. In addition, 10 putative candidate transcripts had an extra exon; 15 a skipped exon and 9<br>with both missing and skipped e site. In addition, 10 putative candidate transcripts had an extra exon; 15 a skipped exon and 9 with both missing and skipped exons relative to the most expressed annotated UJC (Figure 3D). with both missing and skipped exons relative to the most expression relative to the most expression.<br>The most expression of the most expression of the most expression of the most expression of the most expression gene, the putative novel transcript differed from the annotated transcript by an alternative<br>exon in 147 cases (33 extra exons, 86 skipped exons, and 43 with both an extra and skipped<br>exon) (Figure 3E). Details and scripts gene, the putative in the manner, that the manner and annotate manner, the annuality of the supplement exon) (Figure 3E). Details and scripts for this analysis are provided in the Supplementary Methods. exon) (Figure 3E). Details and scripts for this analysis are provided in the Supplementary<br>Methods.<br>This analysis demonstrates that SQANTI-reads is effective in classifying genes as well annotated

exon) (Figure 32). Details and scripts for this analysis of provided in this scripp control.)<br>This analysis demonstrates that SQANTI-reads is effective in classifying genes as well annotated<br>or underannotated and flagging This analy:<br>This analy:<br>alternative  $\frac{1}{2}$ This analysis demonstrates that SQANTI-reads is effective in classifying genes as well annotated<br>or underannotated and flagging putative novel transcript annotations that contain realistic<br>alternative exonic patterns. or under and flagging put and flagging putative exonic patterns.<br>Contained transcript and flagging putations that contain realistical realistic put and contain realistical realistical realistical realistical realistical re alternative exonic patterns.



classifying genes as well annotated or under-annotated and classifying transcripts as putative novel candidate transcripts. B) Number of well-annotated and under-annotated genes according to the decision tree in A. C) The coverage (percent of total reads) vs length (percentage of maximum junctions) for all UJCs in underannotated genes with well covered candidate transcripts. UJCs that meet the thresholds for putative novel transcripts are coloured candidate transcripts. UJCs that meet the thresholds for putative novel transcripts are coloured in green. D&E) Comparison of putative candidate novel transcripts with the most expressed Figure 3: Evaluation of WTC11 PacBio samples for under-annotated genes. A) Decision tree for

annotated transcript in that gene for well-annotated genes with well-covered FSMs (D) and with an FSM detected but without  $\times$  20% of the total reads in the gene(E).



Figure 4. Variation in donors/acceptors. Metrics are only calculated for annotated donor/acceptors with a slip minimum threshold of reads (20 by default). A) Metrics for the classification of donor/acceptor variation. When all metrics for reads align to the annotated donor/acceptor this is classified as a reference match) (Ref Match). When all reads align to the same donor/acceptor location, but this is not the annotated position this is classified as  $Cl$  = 0. When reads align in multiple positions in proximity to an annotated donor/acceptor this is classified as CV + 0. B) Classification of the number (left) and proportion (right) of detected acceptors faceted by technology. C Classification of the number (left) and proportion (right) of detected donors faceted by technology. Figure 4. Variation in donors/acceptors. Metrics are only calculated for annotated donor/acceptors with a  $\overline{a}$ 

## SQANTI-reads metrics for donors/acceptors identify noisy splicing and potentially novel splice-sites

s<br>s<br>s expressed annotated donors/acceptors (Figure 4A). A value CV > 0 indicates variability in the<br>donor/acceptor, with higher CV values indicating more variability. Variability around a splice-<br>junction may be due to weak spli expressed and in the set of values indicating more variability. Variability around a splice-<br>intertion may be due to weak splicing (Wang and Marín 2006) or to technology errors and<br>mapping accuracy, for example due to junc in the variability (CV > 0) in donors and acceptors (Figure 4B). All three technologies identify mapping accuracy, for example due to junction ambiguity (Li 2018). We evaluated these metrics<br>on the WTC11 dataset for reference junctions with at least 10 reads. We found similar patterns on the WTC11 dataset for reference junctions with at least 10 reads. We found similar patterns<br>in the variability (CV > 0) in donors and acceptors (Figure 4B). All three technologies identify<br>donors and acceptors with var on the variability (CV > 0) in donors and acceptors (Figure 4B). All three technologies identify<br>donors and acceptors with variability around the splice site (CV > 0). Donors/acceptors with CV<br>> 0 consistently across the t donors and acceptors with variability around the splice site  $(CV > 0)$ . Donors/acceptors with CV<br>> 0 consistently across the three technologies, are highly suspicious of 'noisy' splicing or a weak<br>splice site and may be wo  $>$  0 consistently across the three technologies, are highly suspicious of 'noisy' splicing or a weak splice site and may be worth follow up (Supplementary Figure 6A and 6B). The SQANTI-reads<br>output file cv.csv identifies the donors/acceptors with CV > 0 making it straightforward to<br>follow up on particular locations with t splice in the matrice of the setting of the matrice of the comput file cv.csv identifies the donors/acceptors with CV > 0 making it straightforward to follow up on particular locations with tools such as Integrative Genomi output the entity included the donorsy-acceptors mini-extracting incomignation and follow up on particular locations with tools such as Integrative Genomics Viewer (IGV)<br>(Robinson et al. 2011).<br>A reference match indicates

follows up the sum of the set of the such as Integrative Center all as Integrative Center and the spl A reference match indicates that the splice signal is strong (Wang and Marín 2006; Dent et al.  $\frac{1}{2}$   $\frac{1}{2}$ A reference match indicates that the split is strong (Tamg and Marin 2007) 2007 in<br>2021). Results for cDNA PacBio and dRNA-seq were similar, with both showing a higher number<br>and proportion of reference match donors/accept and proportion of reference match donors/acceptors compared to cDNA ONT (Figure 4B and C)<br>despite these technologies detecting similar numbers of FSM UJCs (Figure 2D). We compared<br>the FSM UJCs identified by the three techn despite these technologies detecting similar numbers of FSM UJCs (Figure 2D). We compared<br>the FSM UJCs identified by the three technologies (Supplementary Figure 7A). Most of the FSM<br>UJCs are detected by all three technol UJCs are detected by all three technologies ( $n = 19690$ ), with a similar number detected by cDNA PacBio only ( $n = 8110$ ) and cDNA ONT ( $n = 9360$ ) only. We hypothesized that the difference in the number of reference match UJCs are detected by all three technologies (n = 19690), with a similar number detected by<br>cDNA PacBio only (n = 8110) and cDNA ONT (n = 9360) only. We hypothesized that the<br>difference in the number of reference match dono  $CDNA$  PacBio only (n = 8110) and cDNA ONT (n = 9360) only. We hypothesized that the<br>difference in the number of reference match donors/acceptors was potentially due to longer<br>transcripts with more junctions being detected difference in the number of reference match donors/acceptors was potentially due to longer<br>transcripts with more junctions being detected in cDNA PacBio compared to shorter transcripts<br>with fewer junctions in dRNA ONT and transcripts with more junctions being detected in cDNA PacBio compared to shorter transcripts<br>with fewer junctions in dRNA ONT and cDNA ONT. For the FSMs detected only in one<br>technology we plotted the distribution of the n with fewer junctions in dRNA ONT and cDNA ONT. For the FSMs detected only in one technology we plotted the distribution of the number of junctions and confirmed that the cDNA technology we plotted the distribution of  $j$ unctions and confirmed that the c $\mathcal{O}_I$ 

PacBio FSM transcripts had a larger number of junctions compared to dRNA ONT and cDNA<br>ONT (Supplementary Figure 8). This is not surprising given that cDNA PacBio had longer reads<br>than both ONT technologies (Figure 2E).

ONT (Suppriminary Figure 2). This is the complementary given that contribute that conger reads<br>than both ONT technologies (Figure 2E).<br>Donor/acceptors ites may differ from annotated sites due to the presence of novel<br>donor onor/acceptor sites may differ from<br>donors/acceptors. The category CV = 0 id<br>from the annotated donor/acceptor. We l<br>c<br>f donors/acceptors. The category CV = 0 identifies donor/acceptors with no variability, that differ<br>from the annotated donor/acceptor. We evaluated the donors and acceptors with CV = 0 in all<br>the technologies (Supplementary or the annotated donor/acceptor. We evaluated the donors and acceptors with CV = 0 in all<br>the technologies (Supplementary Figure 6E and 6F). Of the 5562 donors and 4474 acceptors<br>with CV = 0 across all technologies, there from the annotation activity acceptor. The annotation are active and acceptors and 4474 acceptors<br>with CV = 0 across all technologies, there were 47 donors and 35 acceptors with CV = 0<br>detected in all three technologies. T with CV = 0 across all technologies, there were 47 donors and 35 acceptors with CV = 0 detected in all three technologies. These had a median distance from the annotated donor or acceptor of 4 nucleotides and 5 nucleotide detected in all three technologies. These had a median distance from the annotated donor or acceptor of 4 nucleotides and 5 nucleotides respectively (Supplementary figure 9). These indicate potential robust detection of alternative splice sites.<br>DISCUSSION/CONCLUSION indicate potential robust detection of alternative splice sites.<br>DISCUSSION/CONCLUSION

SQANTI-reads compares multiple samples using summaries and visualizations of SQANTI3  $\begin{array}{c} \n\frac{1}{2} & \frac{1}{2} \\ \n\frac{1}{2} & \frac{1}{2} \\ \n\end{array}$  $rac{1}{2}$ SALE CONCLUSION CONCLUDED<br>SQANTI-reads compares reategories and sub-categories S<br>C<br>C<br>t Categories and sub-categories that enable the researcher to evaluate the experiment for<br>consistency, and identify any systematic differences between sample groups. The meta-data in<br>the design file determines the sample gro consistency, and identify any systematic differences between sample groups. The meta-data in<br>the design file determines the sample groupings, and once the initial classification and junction<br>files are created, SQANTI-reads the design file determines the sample groupings, and once the initial classification and junction<br>files are created, SQANTI-reads can be invoked multiple times to compare different aspects of<br>the experimental design. the design file design file sample groupings, and once the initial classification and junction<br>files are created, SQANTI-reads can be invoked multiple times to compare different aspects of<br>the experimental design. file experimental design.<br>Files are created are created are invoked multiple times to compare the incompare technology (ONT) the existence of multiple platforms at different price

the experimental design.<br>In Oxford Nanopore Teck<br>points for different num  $\begin{array}{c} 1 \\ 5 \end{array}$  $\begin{array}{c} 1 \\ 5 \\ 6 \end{array}$ In Alternation Contrology (Consension Nanopolius for different numbers of pores (Flongle, MinION, GridION, PromethION) but with the same library protocols means that in a large experiment, samples can be initially evaluate points for anti-ordent numbers of points (Flongle) minion, controlly controlling, and the lines.<br>Same library protocols means that in a large experiment, samples can be initially evaluated at<br>low cost on one of the lower t same library protocols in the library of the low cost on one of the lower throughput platforms (Flongle MinION) and if samples are of sufficient quality, then can be run on higher throughput platforms (GridION, PromethION) sufficient quality, then can be run on higher throughput platforms (GridION, PromethION). sufficient quality, then can be run on higher throughput platforms (GridION, PromethION).

Sample multiplexing and Doerge 2010).<br>Sample multiplexing and Doerge 2010).<br>Sample shows and parally is also good transcripts for further evaluation. As the examples<br>(annotated genes; and putative novel transcripts for fur (Auer and Doerge 2011).<br>SQANTI-reads introduces<br>annotated genes; and presented demonstrate, くこうしょ annotated genes; and putative novel transcripts for further evaluation. As the examples<br>presented demonstrate, SQANTI-reads is flexible and can be customized to examine the impact presented demonstrate, SQANTI-reads is flexible and can be customized to examine the impact<br>of different aspects of the sample meta-data on the metrics for donor/acceptor variation and<br>novel transcript identification. The present aspects of the sample meta-data on the metrics for donor/acceptor variation and<br>novel transcript identification. The output of SQANTI-reads can be easily mined for additional<br>insights and can be used to direct atte novel transcript identification. The output of SQANTI-reads can be easily mined for additional<br>insights and can be used to direct attention and resources to interesting and novel features of<br>IrRNAseq experiments. We expect insights and can be used to direct attention and resources to interesting and novel features of<br>|rRNAseq experiments. We expect SQANTI-reads to become an essential tool for the QC of<br>|multi-sample |rRNA-seq datasets. in the used to direct and the come and essential tool for the QC of<br>insights ample lines are directed.<br>Interesting and resources and resources of the QC of larent to proportion the expect SQANTI-reads to be contain to be considered to the QC of<br>multi-sample lrRNA-seq datasets.<br>Acknowledgements

## multi-sample landscape lines.<br>Acknowledgements.<br>This work we a sex a set of in a set b  $\overline{a}$ Acknowledgements

ヿ ( ド t | | (1R21HG011280-01), the Spanish MICIN (PID2020-119537RB-I00), the European U<br>programme Horizon Europe under the Marie Skłodowska-Curie Actions postdoctc<br>to C.M. (101149931). This work is also supported in part by a grant fr or alles a Correction Currection and the Marie Skłodowska-Curie Actions postdoctoral fello<br>to C.M. (101149931). This work is also supported in part by a grant from the National Car<br>Institute (NCI P01 CA214091), the Univers pro C.M. (101149931). This work is also supported in part by a grant from the National Cancer<br>Institute (NCI P01 CA214091), the University of Florida Department of Molecular Genetics and<br>Microbiology, the University of Flo Institute (NCI P01 CA214091), the University of Florida Department of Molecular Genetics and<br>Microbiology, the University of Florida Genetics Institute, the University of Florida Cancer<br>Center, and the University of Florid Microbiology, the University of Florida Genetics Institute, the University of Florida Cancer<br>Center, and the University of Florida Research Computing Center (<u>WWW.rc.ufl.edu</u>) and the Lati<br>American and Caribbean Scholars a Microbiology, the University of Florida Research Computing Center (www.rc.ufl.edu) and the<br>Center, and the University of Florida Research Computing Center (www.rc.ufl.edu) and the<br>American and Caribbean Scholars award to N American and Caribbean Scholars award to N.K. Part of the computations were performed on<br>the high performance computing cluster Garnatxa at the Institute for Integrative Systems<br>Biology (I2SysBio). I2SysBio is a mixed rese the high performance computing cluster Garnatxa at the Institute for Integrative Systems<br>Biology (I2SysBio). I2SysBio is a mixed research center formed by University of Valencia (UV)<br>and Spanish National Research Council ( the high performance compution is a mixed research center formed by University of Valencia (U<br>and Spanish National Research Council (CSIC). We acknowledge Ashley Myrick for help wi<br>some of the initial CV plots and Knife Ba Biology (I2System, 128). In a minimization controlled parameter, the formed particle and Spanish National Research Council (CSIC). We acknowledge Ashley Myrick for help with some of the initial CV plots and Knife Bankole f some of the initial CV plots and Knife Bankole for the initial coding of the junction hash. Aliso<br>Morse prepared all samples and libraries for the *Drosophila* experiment, ran all initial QC<br>analyses and recalled all the b Some of the information prediction in the initial Component, and all initial QC<br>analyses and recalled all the bases for ONT data. We acknowledge Rolf Renne for his support.<br>Competing interest statement<br>A.C. has received in

Morse prepared an samples and infraries for the Drosophila experiment, ran an initial QC<br>analyses and recalled all the bases for ONT data. We acknowledge Rolf Renne for his supp<br>Competing interest statement<br>A.C. has receiv Competing interest statement<br>A.C. has received in-kind funding from Pacific Biosciences for library preparation and<br>sequencing. A.C. collaborates with Oxford Nanopore in the Marie Skłodowska-Curie Actions<br>Doctoral Network  $\frac{1}{2}$ **Competing interest statement**<br>A.C. has received in-kind funding from Pacific Biosciences for library preparation and Sequencing. A.C. collaborates with Oxford Nanopore in the Marie Skłodowska-Curie A<br>Doctoral Network project LongTREC (101072892).<br>References Doctoral Network project LongTREC (101072892).<br>References Doctoral Network project LongTREC (101072892).

## .<br>. References

S, Schwartz MA, Blaum EM et al. 2024. High-throughput RNA isoform sequencing using<br>programmed cDNA concatenation. Nature Biotechnology 42: 582-586.<br>Amarasinghe SL, Ritchie ME, Gouil Q. 2021. long-read-tools.org: an interac

- 
- programmed cDNA concatenation. Nature Biotechnology 42: 582-586.<br>Stagle SL, Ritchie ME, Gouil Q. 2021. long-read-tools.org: an interactive catalogue of<br>analysis methods for long-read sequencing data. *GigaScience* 10.<br>J. D
- programmed CDNA concatenation. Nature Biotechnology 42: 582-586.<br>inghe SL, Ritchie ME, Gouil Q. 2021. long-read-tools.org: an interactive<br>analysis methods for long-read sequencing data. *GigaScience* 10.<br>., Doerge RW. 2010 amalysis methods for long-read sequencing data. GigaScience 10.<br>Auer PL, Doerge RW. 2010. Statistical Design and Analysis of RNA Sequencing Data. Genetics.<br>185: 405-416.<br>Begik O, Diensthuber G, Liu H, Delgado-Tejedor A, Ko analysis methods for long-read sequencing data. Olydschelle 10.<br>
185: 405-416.<br>
185: 405-416.<br>
186: A05-416.<br>
189: Andren G, Liu H, Delgado-Tejedor A, Kontur C, Niazi AM, Va<br>
Beaudoin JD, Mattick JS et al. 2023. Nano3P-seq Auer PL, Doerge RW. 2010. Statistical Design and Analysis of RNA Sequencing Data. Genetics<br>
185: 405-416.<br>
Begik O, Diensthuber G, Liu H, Delgado-Tejedor A, Kontur C, Niazi AM, Valen E, Giraldez AJ,<br>
Beaudoin JD, Mattick J 185: 405-416.<br>
Begik O, Diensthuber G, Liu H, Delgado-Tejedor A, Kontur C, Niazi AM, Valen E, Giraldez AJ,<br>
Beaudoin JD, Mattick JS et al. 2023. Nano3P-seq: transcriptome-wide analysis of gene<br>
expression and tail dynamics Beaudoin JD, Mattick JS et al. 2023. Nano3P-seq: transcriptome-wide analysis of ger<br>expression and tail dynamics using end-capture nanopore cDNA sequencing. Nat<br>Methods 20: 75-85.<br>Carbonell-Sala S, Perteghella T, Lagarde J
- Example in Symman Present Matter Dependency Present Matematy Presentations, the expression and tail dynamics using end-capture nanopore cDNA sequencing. Nat<br>Methods 20: 75-85.<br>ell-Sala S, Perteghella T, Lagarde J, Nishiyor expression and tail dynamics using end-capture nanopore cDNA sequencing. Nut<br>Methods 20: 75-85.<br>ell-Sala S, Perteghella T, Lagarde J, Nishiyori H, Palumbo E, Arnan C, Takahashi H,<br>Carninci P, Uszczynska-Ratajczak B, Guigó Methods 20: 75-85.<br>ell-Sala S, Perteghell<br>Carninci P, Uszczyns<br>quantitative approac<br>Communications 15:<br>er W, D'hert S, Schu<br>and processing long Carninci P, Uszczynska-Ratajczak B, Guigó R. 2024. CapTrap-seq: a platform-agno:<br>quantitative approach for high-fidelity full-length RNA sequencing. *Nature*<br>*Communications* 15: 5278.<br>De Coster W, D'hert S, Schultz DT, Cr
- quantitative approach for high-fidelity full-length RNA sequencing. Nature<br>
Communications 15: 5278.<br>
De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing<br>
and processing long-read seq
- Gommunicative approach for high-haenty fun length RNA sequencing. Nature<br>Communications 15: 5278.<br>Ler W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: v<br>and processing long-read sequencing data. *Bioinf* and processing long-read sequencing data. *Bioinformatics* 34: 2666-2669.<br>, Singh S, Mukherjee S, Mishra S, Sarwade RD, Shamaya N, Loo KP, Harriso<br>Sureshkumar S, Powell D et al. 2021. Quantifying splice-site usage: a simpl and processing long read sequencing data. Bioinformatics 34: 2000-2005.<br>
Singh S, Mukherjee S, Mishra S, Sarwade RD, Shamaya N, Loo KP, Harriso<br>
Sureshkumar S, Powell D et al. 2021. Quantifying splice-site usage: a simpl<br> Sureshkumar S, Powell D et al. 2021. Quantifying splice-site usage: a simple ye<br>approach to analyze splicing. NAR Genomics and Bioinformatics 3.<br>Fukasawa Y, Ermini L, Wang H, Carty K, Cheung MS. 2020. LongQC: A Quality Con
- Fukasawa Y, Ermini L, Wang H, Carty K, Cheung MS. 2020. LongQC: A Quality Control Tool for<br>Third Generation Sequencing Long Read Data. G3 (Bethesda) 10: 1193-1196.
- Garimella K et al. 2022. Transcriptome variation in human tissues revealed by long-read<br>sequencing. Nature 608: 353-359. approach to analyze splicing. MAN Genomics and Biomformatics 3.<br>wa Y, Ermini L, Wang H, Carty K, Cheung MS. 2020. LongQC: A Qual<br>Third Generation Sequencing Long Read Data. *G3 (Bethesda)* 10: 1<br>DA, Garborcauskas G, Hoffma Functionary Termini L, Thing H, Termin, Wang H, Carty H, Check Carty Control Third Generation Sequencing Long Read Data. G3 (Bethesda) 10: 1193-1196.<br>Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X Third Generation Sequencing Long Read Data. OS (*Dethesdu)* 10: 1153-1130.<br>DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, I<br>Garimella K et al. 2022. Transcriptome variation in human tissues rev Garimella K et al. 2022. Transcriptome variation in human tissues revealed by long-read<br>sequencing. Nature 608: 353-359.<br>Holmqvist I, Bäckerholm A, Tian Y, Xie G, Thorell K, Tang KW. 2021. FLAME: long-read<br>bioinformatics t
- 
- Sequencing. *Nature* 608: 353-359.<br>Sequencing. *Nature* 608: 353-359.<br>Vist I, Bäckerholm A, Tian Y, Xie G, Thorell K, Tang KW. 2021. FLAME: long-read<br>bioinformatics tool for comprehensive spliceome characterization. *Rna* sequencing. *Nature* **008**: 353-359.<br>
Vist I, Bäckerholm A, Tian Y, Xie G, T<br>
bioinformatics tool for comprehen<br>
r A, Hu W, Zhang B, Narykov O, Die<br>
TA, Fedrigo O et al. 2024. Single-ce<br>
specialized splicing patterns in de bioinformatics tool for comprehensive spliceome characterization. Rna 27: 112<br>Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, Balacco J, Ndhlovu LC<br>TA, Fedrigo O et al. 2024. Single-cell long-read sequencingbioinformatics tool for comprehensive spitecome characterization. *Rna 27*: 1127-1139.<br>
r A, Hu W, Zhang B, Narykov O, Diekhans M, Marrocco J, Balacco J, Ndhlovu LC, Milner<br>
TA, Fedrigo O et al. 2024. Single-cell long-read TA, Fedrigo O et al. 2024. Single-cell long-read sequencing-based mapping reveals<br>specialized splicing patterns in developing and adult mouse and human brain. Nature<br>Neuroscience 27: 1051-1063.<br>Leger A, Leonardi T. 2019. p specialized splicing patterns in developing and adult mouse and human brain. Natu<br>Neuroscience 27: 1051-1063.<br>, Leonardi T. 2019. pycoQC, interactive quality control for Oxford Nanopore Seque<br>Journal of Open Source Softwar
- 
- Specialized splicing patterns in developing and addit mouse and human brain. Nature<br> *Neuroscience* 27: 1051-1063.<br>
, Leonardi T. 2019. pycoQC, interactive quality control for Oxford Nanopore Sequencir<br> *Journal of Open So* Neuroscience 27: 1051-1005.<br>
, Leonardi T. 2019. pycoQC, in<br>
Journal of Open Source Softwi<br>
I, Alles J, Karaiskos N, Ayoub S<br>
sequencing reveals principles<br>886.<br>
18. Minimap2: pairwise alignr Journal of Open Source Software 4: 1236.<br>Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. 2019. FLAM-seq: full-length mRNA<br>sequencing reveals principles of poly(A) tail length control. Nature Methods 16: 879-<br>886.<br>Li Journal of Open Source Software 4: 1236.<br>
I, Alles J, Karaiskos N, Ayoub S, Rajewsky I<br>
sequencing reveals principles of poly(A) ta<br>
886.<br>
18. Minimap2: pairwise alignment for nuc<br>
3100.<br>
ndsaker B, Wysoker A, Fennell T, R Sequencing reveals principles of poly(A) tail length control. Nature Methods 16:<br>886.<br>Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:<br>3100.<br>Li H, Handsaker B, Wysoker A, Fennell T, R
- sequencing reveals principles of poly(A) tail length control. Nature Methods 16: 879<br>886.<br>18. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094<br>3100.<br>ndsaker B, Wysoker A, Fennell T, Ruan J, 18. N<br>3100<br>ndsa<br>The S<br>d M,<br>Herw
- 
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 200<br>Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 200<br>The Sequence Alignment/Map format an ndsak<br>The Se<br>d M, v<br>Herwig<br>analys The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.<br>Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Börno S, Caiment F, Vingron M,<br>Herwig R. 2023. IsoTools: a flexible workflow for lon The Sequence Alignment/Map format and SAMtools. Biologymuttes 25. 2070 2075.<br>d M, van den Beucken T, Timmermann B, Hochradel M, Börno S, Caiment F, Vingron<br>Herwig R. 2023. IsoTools: a flexible workflow for long-read transc Herwig R. 2023. IsoTools: a flexible workflow for long-read transcriptome sequencing<br>analysis. *Bioinformatics* 39. Hering A. 2023. Isotophenics: A flexible workflow for long-reading groups analysis. Bioinformatics: 39. analysis. *Bioinformatics* 39.

- 
- Liu P, Sanalkumar R, Bresnick EH, Keleş S, Dewey CN. 2016. Integrative analysis with ChIP-seq<br>advances the limits of transcript quantification from RNA-seq. *Genome Res* 26: 1124-<br>1133.<br>Mahmoud M, Huang Y, Garimella K, Aud advances the limits of transcript quantification from RNA-seq. Genome Res 20: 1124<br>1133.<br>Jud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S,<br>Pionzio A, Schatz MC et al. 2024. Utility of long-read oud M,<br>Pionzio<br>*Comm*<br>. 2023<br>v, Titus<br>Morse
- 
- Pionzio A, Schatz MC et al. 2024. Utility of long-read sequencing for All of Us. Natu<br>Communications 15: 837.<br>Marx V. 2023. Method of the year: long-read sequencing. Nature Methods 20: 6-11.<br>Nanni A, Titus-McQuillan J, Ban Pionzio A, Schatz MC et al. 2024. Utility of long-read sequencing for All of Us. Matare<br>Communications 15: 837.<br>2023. Method of the year: long-read sequencing. Nature Methods 20: 6-11.<br>1, Titus-McQuillan J, Bankole KS, Par Communications 13: 837.<br>2023. Method of the yea<br>1, Titus-McQuillan J, Banko<br>Morse Alison M, Rogers R<br>quantify alternative splicir<br>in JRB, Concannon P, Tard<br>Transcript Events To Impre Marx V. 2025. Method of the year: long-read sequencing. Mature Methods 20: 0-11.<br>
Nanni A, Titus-McQuillan J, Bankole KS, Pardo-Palacios F, Signor S, Vlaho S, Moskaler<br>
Morse Alison M, Rogers RL, Conesa A et al. 2024. Nucl
- Morse Alison M, Rogers RL, Conesa A et al. 2024. Nucleotide-level distance metrics<br>quantify alternative splicing implemented in TranD. Nucleic Acids Research 52: e28-<br>Newman JRB, Concannon P, Tardaguila M, Conesa A, McInty quantify alternative splicing implemented in TranD. Nucleic Acids Research 52: e28-e2<br>
In JRB, Concannon P, Tardaguila M, Conesa A, McIntyre LM. 2018. Event Analysis: Usin<br>
Transcript Events To Improve Estimates of Abundan
- quantify alternative splicing implemented in TranD. Nucleic Acids Research 52: e28-e28.<br>In JRB, Concannon P, Tardaguila M, Conesa A, McIntyre LM. 2018. Event Analysis: Using<br>Transcript Events To Improve Estimates of Abunda Transcript Events To Improve Estimates of Abundance in RNA-seq Data. *G3*<br>Genes | Genomes | Genetics 8: 2923-2940.<br>Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, Matthews BB,<br>Millburn G, Transcript Events To Improve Estimates of Abundance in RNA-seq Data. OS<br>Genes / Genomes / Genetics 8: 2923-2940.<br>Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, M<br>Millburn G, dos Santos G, Tabone Genes, Genomes, Genetics 8: 2923-2940.<br>Colak A, Marygold SJ, Antonazzo G, Attrill<br>Millburn G, dos Santos G, Tabone CJ et al.<br>genes and genomes database. *Genetics* 2<br>Palacios FJ, Arzalluz-Luque A, Kondratova<br>Estevan-Morió
- Millburn G, dos Santos G, Tabone CJ et al. 2024. FlyBase: updates to the Drosophila<br>genes and genomes database. *Genetics* 227.<br>Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R,<br>Estev Milleum Milleum Cymrech Christin Control of the Dreis Agency Sensisting Pressure Constantine Procepting<br>Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R,<br>Estevan-Morió E, Liu T, Nanni A, Mc Palacios FJ, Arzalluz-Luque A, Kondratova L, Salacios FJ, Arzalluz-Luque A, Kondratova L, Salacios FJ, Arzalluz-Luque A, Kondratova L, Salacios **21**: 793-797.<br>Palacios FJ, Wang D, Reese F, Diekhans M, Calacios FJ, Wang D, Estevan-Morió E, Liu T, Nanni A, McIntyre L et al. 2024a. SQANTI3: curation of long<br>transcriptomes for accurate identification of known and novel isoforms. Nature<br>Methods 21: 793-797.<br>Pardo-Palacios FJ, Wang D, Reese F, Di
- transcriptomes for accurate identification of known and novel isoforms. Nature<br>Methods 21: 793-797.<br>Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De<br>María M, Adams MS, Balderrama-Guti Transcriptomes for accurate identification of known and novel isoforms. Nature<br>Methods 21: 793-797.<br>Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland<br>María M, Adams MS, Balderrama-Gutierrez G Methods 21: 753-797.<br>Palacios FJ, Wang D, Re<br>María M, Adams MS, E<br>long-read RNA-seq me<br>*Methods* 21: 1349-136<br>ry A, Zhang P, Jops C, \<br>2024. Developmental i María M, Adams MS, Balderrama-Gutierrez G et al. 2024b. Systematic assessment of<br>long-read RNA-seq methods for transcript identification and quantification. Nature<br>*Methods* 21: 1349-1363.<br>Patowary A, Zhang P, Jops C, Vuon
- Maria M, Maria M, Januariana Fantaria Correla Mercina (preminio attenum attenum attenum)<br>Methods 21: 1349-1363.<br>ry A, Zhang P, Jops C, Vuong CK, Ge X, Hou K, Kim M, Gong N, Margolis M, Vo D et al.<br>2024. Developmental isofo long-read RNA-seq methods for transcript identification and quantification. Nuture<br>Methods 21: 1349-1363.<br>ry A, Zhang P, Jops C, Vuong CK, Ge X, Hou K, Kim M, Gong N, Margolis M, Vo D et a<br>2024. Developmental isoform diver Wethods 21: 1349-1363.<br>
ry A, Zhang P, Jops C, Vuo<br>
2024. Developmental iso<br>
neuropsychiatric risk meo<br>
AR, Hall IM. 2010. BEDT<br>
Bioinformatics 26: 841-84<br>
ND JT, Thorvaldsdóttir H, \ Patricia, P., 2010, 2024, Developmental isoform diversity in the human neocortex informs<br>neuropsychiatric risk mechanisms. Science 384: eadh7688.<br>Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for compa
- neuropsychiatric risk mechanisms. *Science* 384: eadh7688.<br>1 AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing ge<br>*Bioinformatics* 26: 841-842.<br>2011. Thorvaldsdóttir H, Winckler W, Guttman M, Lander neuropsychiatric risk mechanisms. Science 384: eadh7688.<br>
1 AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for<br> *Bioinformatics* 26: 841-842.<br>
2011. Thorvaldsdóttir H, Winckler W, Guttman M, Lander Es<br>
Integrat
- 
- Bioinformatics 26: 841-842.<br>
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011.<br>
Integrative genomics viewer. Nat Biotechnol 29: 24-26.<br>
Tardaguila M, de la Fuente L, Marti C, Perei Biompormaties 20: 841-842.<br>
Signal JT, Thorvaldsdóttir H, Win<br>
Integrative genomics viewer<br>
Sila M, de la Fuente L, Marti<br>
Mellado M, Macchietto M, V<br>
of long-read transcript seque<br>
identification and quantifica Integrative genomics viewer. Nat Biotechnol 29: 24-26.<br>Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M,<br>Mellado M, Macchietto M, Verheggen K et al. 2018. SQANTI: extensive charac Integrative genomics viewer. Mat Biotechnol 29: 24-26.<br>
Illa M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios<br>
Mellado M, Macchietto M, Verheggen K et al. 2018. SQ,<br>
of long-read transcript sequences for quality contr Mellado M, Macchietto M, Verheggen K et al. 2018. SQANTI: extensive characteriz<br>
of long-read transcript sequences for quality control in full-length transcriptome<br>
identification and quantification. Genome Res 28: 396-411 of long-read transcript sequences for quality control in full-length transcriptome<br>identification and quantification. *Genome Res* 28: 396-411.<br>CEL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, Hernandez C
- 
- of longitude transcript sequences for quality sectors. 396-411.<br>
SEL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, Hernandez C<br>
Genomics in the long-read sequencing era. Trends in Genetics 39: 649-671.<br>
R, identification and quantification. Genome Res 28: 396-411.<br>
EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, The<br>
Genomics in the long-read sequencing era. *Trends in Geneti*<br>
R, Palmer T, Byrne A, Cole C, Schmitz RJ Genomics in the long-read sequencing era. Trends in Genetics 39: 649-671.<br>Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. 2018. Improving<br>nanopore read accuracy with the R2C2 method enables the seque Genomics in the long-read sequencing era. Trends in Genetics 39: 649-671.<br>
Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. 2018. Improving<br>
nanopore read accuracy with the R2C2 method enables the seq The R, Palmer T, Byrne C, Schmitz R, Premium, Premium, Premium Premium, Premiu multiplexed full-length single-cell cDNA. Proceedings of the National Academy of
- Sciences 115: 9726-9731.<br>Sciences 115: 9726-9731.<br>M, Marín A. 2006. Characterization and prediction of alternative splice sites. *Gene*<br>219-227. Sciences **115**: 9720-9731.<br>M, Marín A. 2006. Characte<br>219-227. Wang M, Marin A. 2006. Characterization and prediction or alternative splice sites. Gene 366.<br>219-227.

W, Williams B, Trout D et al. 2020. A technology-agnostic long-read analysis pipeline fo<br>transcriptome discovery and quantification. *bioRxiv* doi:10.1101/672931: 672931.<br>Transcriptome discovery and quantification. *bioRxi*  $\frac{1}{2}$ <br>transcriptome discovery and quantification. *bioRxiv* doi:10.1101/672931: 672931.<br> $\frac{1}{2}$ transcriptome discovery and quantification. bioRxiv doi:10.1101/672931: 672931.