Overview of how sequencing data were analyzed

Summary

- All public data for Invitrogen™ Collibri™ kits were produced using public, freely available tools.
- FASTQ files containing results generated from Invitrogen™ Collibri™ DNA or Stranded RNA library preparation kits are available for download from **Connect**, our cloud-based platform.

Introduction

Next-generation sequencing (NGS) is becoming a key approach for investigating the molecular basis of diseases. Because of its sensitivity and specificity, NGS has been replacing legacy technologies for exploring diseases at the molecular level. However, NGS has faced some unique challenges due to the rate of data production outpacing that predicted by Moore's law [1]. The rate of improvements in DNA sequencing is instead predicted on the basis of the Carlson curve—a term coined by *The Economist* in an article in 2006 [1]. The National Human Genome Research Institute (NHGRI) tracks the DNA sequencing cost per human genome and also per nucleotide base, and updates these statistics yearly [2]. NHGRI also graphs the prediction of Moore's law for reference.

With the advent of the Illumina™ NovaSeq™ 6000 sequencer, large-scale human genome studies are available to most researchers. The handling and analysis of the massive amounts of data produced pose formidable challenges and raise important questions about how to ensure transparency of data analysis procedures when researchers share data with the scientific community.

The following bioinformatic pipelines were used to create the data presented by Thermo Fisher Scientific in technical notes that are associated with Collibri library preparation kits. The pipeline for analysis of Collibri RNA libraries is available in user-friendly formats from Genialis™ software. Please go to **thermofisher.com/collibrianalysis** to request a demo.

DNA sequencing analysis

The DNA sequencing data from the NovaSeq 6000 platform were analyzed using multiple software packages for base calling, quality control (QC), and analysis. All libraries were demultiplexed using Illumina™ bcl2fastq2 Conversion Software v2.19 [3]. Adapters were trimmed from the reads using AdapterRemoval v2.2.2. Data QC was performed multiple times throughout the pipeline using FastQC v0.11.5 software [4]. The downsizing of the samples to equal read counts was accomplished with SeqKit v0.7.0. Sequence alignment was carried out using BWA v0.7.15. Sorting of reads to understand sequencing or gene depth was accomplished with Sambamba v0.6.6. Alignment rates and GC bias statistics were completed with Picard v2.7.1. Insert size and alignment statistics were completed using Qualimap v2.2.1. Duplicates and chimeric reads were analyzed using SAMtools v1.4. The items in Table 1 are the specific terms used in the programs for DNA analysis.

RNA sequencing analysis

RNA sequencing was conducted on multiple RNA samples. All libraries were sequenced on an Illumina™ HiSeq™ 4000 sequencing system. The resulting libraries were demultiplexed using bcl2fastq2 Conversion Software v2.19 and were trimmed using the BBDuk trimming and filtering tool from the BBmap v.37.90 package.

SeqKit v0.7.0 was used for downsampling. The data were aligned using STAR v2.5.3a, with chimeric read detection system enabled. The alignment was completed using the hg19 reference genome from UCSC [5] and supplemented with the sequences of the ERCC spike-in control. Sorting of the aligned reads was accomplished with Sambamba v0.6.6. Quality measurements were completed and read counts per gene were determined using the QoRTs v1.1.8 program. Read distributions across introns and exons were characterized using the read_distribution.py program from RSeQC package v2.6.2. The differential gene expression was measured using DESeq2 v1.16. Fragments per kilobase per million reads mapped (FPKM) evaluation for genes was accomplished using StringTie v1.3.3. The genome annotations for the analyses were taken from GENCODE [6]. The main file containing comprehensive gene annotation on the reference chromosomes only was used. For ribosomal RNA quantitation, an extended annotation file was used that incorporated additional scaffolds [7]. The specific terms used for the RNA-sequencing analysis for each of the programs are shown in Table 2.

Table 1. DNA sequencing analysis program parameters.

Workflow	Program	Parameters
Data QC	FastQC	-k 5
Trimming of reads	AdapterRemoval	--maxns 1 --minlength 20 --trimqualities --minquality 30 --adapter1 {fw_adapter_sequence} --adapter2 {rv_adapter_sequence} --gzip --barcode-mm 2 --barcode-mm-r1 1 --barcode-mm-r2 1
Downsampling	SeqKit	sample -s 11 -j {threads} --two-pass -n ~300 mln
Alignment	BWA-MEM | SAMtools	-t {threads} {reference} -r 1.0 -k 19 -M -B 6 -v 1 | samtools view -@ 3 -bS - > aligned.bam
Sorting	Sambamba	sort -t {threads} -p -o sorted.bam
Alignment rate	Picard	CollectAlignmentSummaryMetrics
GC bias	Picard	CollectGcBiasMetrics
Insert size, alignment statistics	Qualimap	bamqc -nw 5000 -ip
Sequence or gene depth	Sambamba | sed	depth region -t {threads} -L {input} {bed} | sed '/^#/ d' > {output}
Chimeric reads	SAMtools	samtools view {input} | grep 'SA:' | wc -l
Duplicates	SAMtools	samtools rmdup {input} {dedupe.bam} 2>&1 | tee log

Table 2. RNA sequencing analysis program parameters.

Workflow	Program	Parameters
Trimming of reads	BBDuk from BBmap v.37.90 package	ktrim=r k=23 mink=11 hdist=1 minlength=50 maxns=1 tpe tbo qtrim=r trimq=15
Downsampling	SeqKit v0.7.0	sample -s 11 -j {threads} --two-pass -n ~47 mln
Alignment: aligned against hg19 supplemented with sequences of ERCC spike-in control	STAR v2.5.3a	--chimSegmentMin 20 --outSAMstrandField intronMotif --outSAMattributes All --outSAMtype BAM Unsorted
Read counts per gene; read distribution across gene biotypes; strandedness and other quality measurements	QoRTs v1.1.8*	--maxPhredScore 126 --noGzipOutput --maxReadLength 151 --minMAPQ 50 --addFunctions annotated SpliceExonCounts, FPKM
FPKM values per gene	StringTie v1.3.3	-e -m 50
Read distributions across introns and exons	read_distribution.py from RSeQC package v2.6.2	Default options
Differential expression	DESeq2 v1.16	Default options

* For rRNA amount evaluation in QoRTs analysis, the extended annotation version covering additional scaffolds was used. The "--keep MultiMapped" flag was also added.

Genialis software user interface

The Collibri Stranded RNA Library Prep Kits for high-throughput Illumina systems are verified for whole-transcriptome and mRNA analysis using Genialis software. Collibri Stranded RNA Library Prep Kits make it possible to capture a faithful representation of virtually all types of RNA in a sample. Genialis software empowers analysis and interpretation of these rich and diverse data.

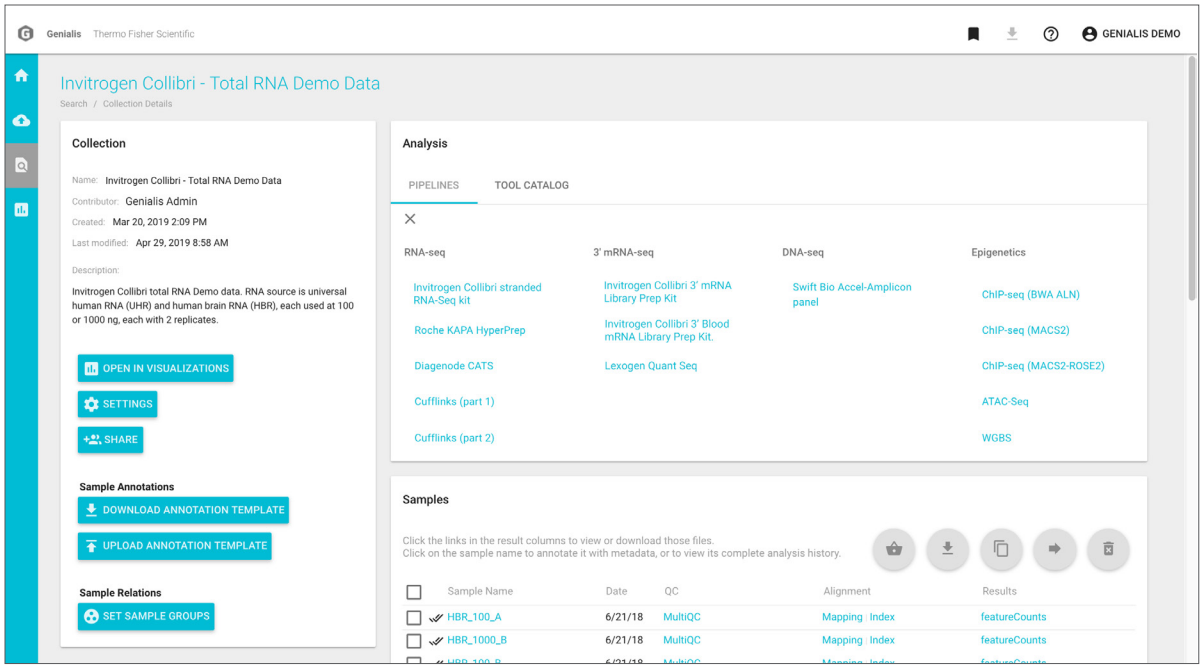

Figure 1. Data analysis using Genialis software. Data are organized into samples and collections, each with a layer of metadata. Preconfigured pipelines can be run with a click of a button, but can also be modified and extended with a selection of bioinformatic tools.

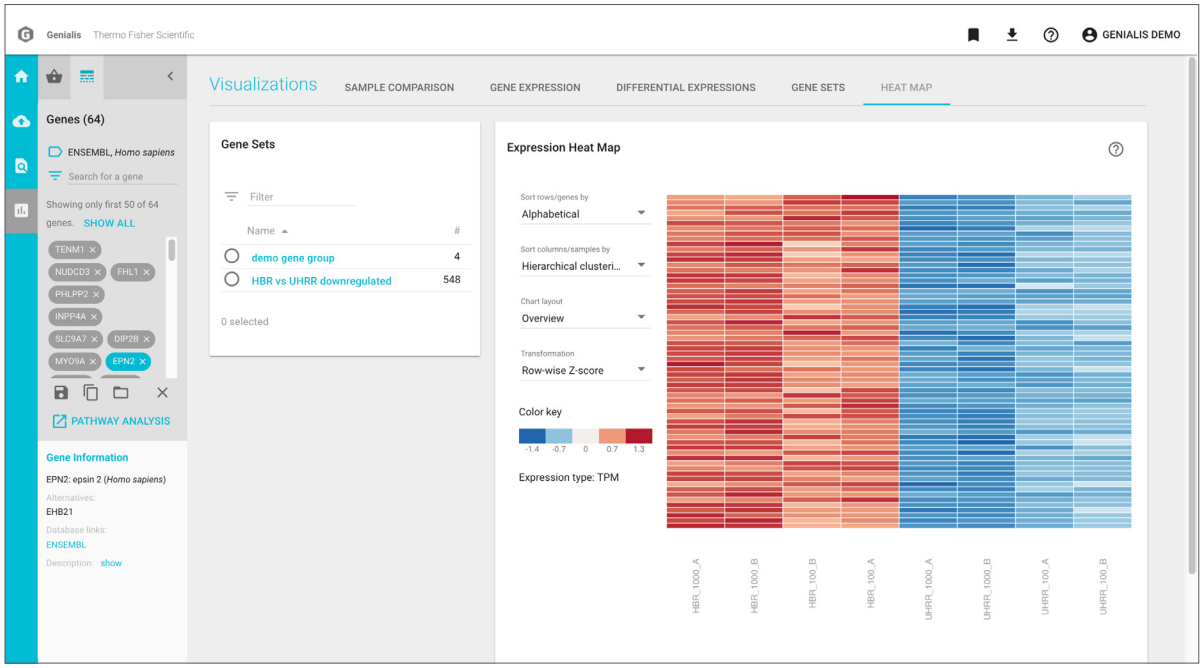

Figure 2. Visualization of workflow modules. Visual modules are arranged into workflows that examine experimental design, gene expression, differentially expressed genes, and more. End users can autonomously explore expression data in real time—freeing bioinformaticians from constant requests and allowing them to focus on the science.

Public data sets

FASTQ files are available for download from **Connect** for libraries generated using Collibri DNA library preparation kits from Coriell NA12878, FFPE, and bacterial samples. FASTQ files are available for download from Connect for libraries generated using Collibri Stranded RNA library preparation kits from human brain (HBr) and Universal Human Reference RNA (UHRR) samples. RNA-Seq FASTQ files are available with poly(A) reads retained or trimmed out for comparison among analysis pipelines.

References

1. Life 2.0. **https://www.economist.com/node/7854314** (accessed July 2019).
2. The cost of sequencing a human genome. **https://www.genome.gov/sequencingcosts/** (accessed July 2019).
3. bcl2fastq2 Conversion v.2.19 User Guide. Document No. 15051736 v02.
4. Babraham Bioinformatics. **https://www.bioinformatics.babraham.ac.uk/projects/fastqc/** (accessed July 2019).
5. hg19 reference genome. **http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips** (accessed July 2019).
6. GENCODE release 19. **https://www.gencodegenes.org/releases/19.html** (accessed July 2019).
7. Annotation file for ribosomal RNA quantitation: **ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.chr_patch_hapl_scaff.annotation.gtf.gz**

Ordering information

Product		Quantity	Cat. No.
DNA-Seq kits for Illumina systems			
Collibri ES DNA Library Prep Kits	with CD Indexes	24 preps	A38605024
	with CD Indexes	96 preps	A38607096
	with UD Indexes, Set A (1-24)	24 preps	A38606024
	with UD Indexes, Set B (25-48)	24 preps	A43605024
	with UD Indexes, Set C (49-72)	24 preps	A43606024
	with UD Indexes, Set D (73-96)	24 preps	A43607024
Collibri PCR-Free ES DNA Library Prep Kits	with CD Indexes	24 preps	A38545024
	with CD Indexes	96 preps	A38603096
	with UD Indexes, Set A (1-24)	24 preps	A38602024
	with UD Indexes, Set B (25-48)	24 preps	A43602024
	with UD Indexes, Set C (49-72)	24 preps	A43603024
	with UD Indexes, Set D (73-96)	24 preps	A43604024
Collibri PS DNA Library Prep Kits	with CD Indexes	24 preps	A38612024
	with CD Indexes	96 preps	A38614096
	with UD Indexes, Set A (1-24)	24 preps	A38613024
	with UD Indexes, Set B (25-48)	24 preps	A43611024
	with UD Indexes, Set C (49-72)	24 preps	A43612024
	with UD Indexes, Set D (73-96)	24 preps	A43613024
	with UD Indexes, Set A-D (1-96)	96 preps	A38614196
Collibri PCR-Free PS DNA Library Prep Kits	with CD Indexes	24 preps	A38608024
	with CD Indexes	96 preps	A38610096
	with UD Indexes, Set A (1-24)	24 preps	A38609024
	with UD Indexes, Set B (25-48)	24 preps	A43608024
	with UD Indexes, Set C (49-72)	24 preps	A43609024
	with UD Indexes, Set D (73-96)	24 preps	A43610024
	with UD Indexes, Set A-D (1-96)	96 preps	A38615196

CD = combinatorial dual, UD = unique dual

invitrogen

Ordering information (continued)

Product	Quantity	Cat. No.
RNA-Seq kits for Illumina systems		
Collibri Stranded RNA Library Prep Kit for Illumina Systems	24 preps	A38994024
	96 preps	A38994096
Collibri Stranded RNA Library Prep Kit for Illumina Systems with H/M/R rRNA Depletion Kit	24 preps	A39003024
	96 preps	A39003096
ERCC RNA Spike-In Mix	1 kit	4456740
ERCC ExFold RNA Spike-In Mixes	1 kit	4456739
Library quantification		
Collibri Library Quantification Kit	100 rxns	A38524100
	500 rxns	A38524500
Qubit 4 Fluorometer, with WiFi	1 fluorometer	Q33238
Qubit 4 NGS Starter Kit, with WiFi	1 kit	Q33240
Library amplification		
Collibri Library Amplification Master Mix	50 rxns	A38539050
	250 rxns	A38539250
Collibri Library Amplification Master Mix with Primer Mix	50 rxns	A38540050
	250 rxns	A38540250

H/M/R = human/mouse/rat

Find out more at **thermofisher.com/collibri**

ThermoFisher SCIENTIFIC