TGC AT YOUR SERVICE

Taking your research to the next generation

TechnionGenomeCenter

TECHNION
Israel Institute of Technology

1. TGC At your service
2. Applications of Next Generation Sequencing
3. Experimental design
4. TGC workflow
5. Sample preparation
6. Illumina sequencing technology
7. Bioinformatics
8. State of the art equipment

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

1.

TECHNION GENOME CENTER AT YOUR SERVICE

Your groundbreaking research is pushed forward by our ambitious center – the Technion Genome Center.

The ongoing exponential advancement of sequencing technologies has ushered in a new and exciting era for biological research. This quickly evolving technology provides researchers with new tools for answering biological questions in a largescale, high-throughput manner.

Gene expression is now detected for *all genes* in hundreds, or even thousands, of samples at a single-cell resolution. Whole genomes are mapped for entire cohorts of individuals to pinpoint individual disease-linked or trait linked polymorphisms. The possibilities are endless for designing experiments using this "next generation" technology! For this you need a service you can trust with your project.

Since 2009, the Technion Genome Center (TGC) has poised itself at the forefront of sequencing technology by continuously upgrading its state of the art technology.

Located in the Technion's new Emerson Life Sciences Building on the main campus, the TGC team is at your service with dedicated bioinformaticians and molecular biology specialists. Our team has a reputation of providing researchers with expert, personalized service from the beginning stages of experimental design, library preparation, and sequencing, to bioinformatic analysis.

At the TGC we take pride in providing researchers with the tools and guidance necessary to make each project a success.

Technion Genome Center Team

Our team includes chief scientists, lab technicians and professional bioinformaticians who are always happy to provide you with the best service possible.

We're here to answer your questions:

TECHNION GENOME CENTER (TGC)
Emerson Building for Life Science, Technion
Haifa 32000 Israel
Tel: 972-4-829-5168
Fax: 972-4-8295131

tgc@tx.technion.ac.il
http://tgc.net.technion.ac.il

Key Benefits of Sequencing at the TGC

- The Technion Genome Center was the first lab in Israel to offer Next Generation Sequencing services using the revolutionary Illumina HiSeq 2500 and MiSeq platforms.
- The Full Package! Our services include: expert consultations on experimental design, sample preparation, high-throughput sequencing, and bioinformatic analysis.
- The TGC is committed to supporting researchers from sample preparation to data analysis in order to help increase productivity and strengthen understanding and use of Next-Generation Sequencing techniques.
- Direct, personal interaction with the bioinformatician committed to each project.

TGC Commitments:

- To have the most up to-date technology
- To provide researchers with fast, reliable, and high quality service

DID YOU KNOW?

The Technion Genome Center is constantly working together with researchers to establish new sequencing applications.

The CEL-Seq protocol for multiplexed RNA-Seq of individual cells (Hashimshony et al. *Cell Reports*, 2012) was developed in the Yanai lab at the Technion and is now available as a service at the Technion Genome Center (see page p8)

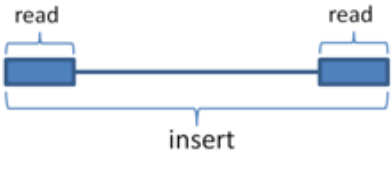
We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Basic concepts in high-throughput sequencing

The following are basic definitions important for high-throughput sequencing:

Insert:
The DNA fragment that is used for sequencing.



Read:
The part of the insert that is sequenced.

Single Read (SR):
A sequencing method in which the insert is sequenced from one end only.

Paired End (PE):
A sequencing method in which the insert is sequenced from both ends.

Flow Cell:
A small glass chip on which DNA fragments are bound and sequenced. The flow cell is covered by DNA probes to which adaptor-ligated DNA fragments hybridize for sequencing.

Lane:
Each flow cell consists of physically separated channels called "lanes." MiSeq flow cells contain one lane each, whereas HiSeq 2500 flow cells have either eight or two lanes, depending on the selected sequencing mode (high-throughput or Rapid, respectively). In both modes all lanes of the flow cell are sequenced simultaneously.

Multiplexing/Demultiplexing:
Sequencing multiple samples on the same lane is called "multiplexing". The bioinformatic separation of reads from multiple samples that were sequenced together on one lane is called "demultiplexing" and is done by a script that recognizes the index of each read and compares it to the known indices of each sample.

Pipeline:
A series of computational commands for bioinformatic analyses.

APPLICATIONS OF NEXT GENERATION SEQUENCING

2.

High-throughput sequencing applications can be divided into two main categories: “reading” and “counting.” In reading applications the focus of the experiment is the sequence itself, such as identifying genomic variants or assembling the sequence of an unknown genome. Counting applications are typically used for quantification of various reads, which can then be compared, such as in gene expression level comparisons.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

High-Throughput Sequencing – common applications

Resequencing:
Sequencing of a whole genome, mapping it to a reference genome and finding variants such as SNPs, insertions, and deletions. The required coverage is usually approximately 30X or more, or according to the requirements of the specific project. The length and type of reads depend on the desired coverage and genome size.

***De Novo* Sequencing:**
Sequencing of DNA whose genome is not previously known. Contigs and scaffolds are generated in the analysis. This application requires long, paired-end reads and high coverage.

Exome Sequencing:
An exon-enriched DNA library is sequenced and reads are mapped to the genome. Variants (SNPs and small indels) are found using bioinformatics tools and compared to databases of known genetic variation. In addition, samples are compared and positions in the genome that show the required combinations are identified. This application requires 100 bp paired-end reads in order to obtain proper coverage.

mRNA-Seq – Gene Expression Analysis:
mRNA is isolated from total RNA, fragmented, and transcribed to cDNA. A sequencing library is then created by ligation of adaptors containing unique barcodes and amplification of adaptor-bound fragments. The sequenced reads are mapped to the genome and normalized, thereby enabling comparison of gene expression levels between samples. In most cases this application requires 50bp single-read sequencing, unless otherwise specified as in projects analyzing splice junctions/ splice variants.

ChIP-Seq:
DNA fragments, usually enriched by specific protein binding sites, are sequenced alongside control DNA ("input DNA"). Enriched regions ("peaks"/"islands") in the genome are identified by comparing the ChIP and input DNA samples.

Small RNA-Seq:
Small RNA sequencing is a powerful application, enabling the discovery and profiling of small RNA and microRNA sequences. This application requires only a short single-read run.

Single-cell RNA-Seq at the TGC

High-throughput sequencing has become an invaluable tool for conducting detailed gene expression analyses, yet the requirement of relatively high RNA starting amounts has posed a challenge for single cell analyses.

The Yanai lab at the Technion has developed a single-cell transcriptomics protocol that overcomes this limitation by uniquely barcoding each sample, and then pooling multiple samples in order to reach the required input amount for mRNA amplification via *in vitro* transcription (Hashimshony et al. *Cell Reports*, 2012). CEL-Seq gives highly reproducible, linear, and sensitive results, all at reduced prices thanks to multiplexing. The robust transcriptome quantification enabled by CEL-Seq is overwhelmingly useful for transcriptomic analyses, such as dissecting complex tissues containing populations of diverse cell types.

The Technion Genome Center provides a service for researchers wishing to apply the CEL-Seq technology. Amplified RNA prepared by the researcher is submitted to the TGC for Illumina library preparation and sequencing.

The TGC also provides bioinformatics services for the resulting gene expression data. This enables researchers to profile tens of samples (each a single cell, tissue sample, embryo, etc.), if not more, from a single Illumina flow cell lane, thereby unlocking the power of RNA-Seq for transcriptomic analysis on a single-cell level.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

CEL-Seq Protocol

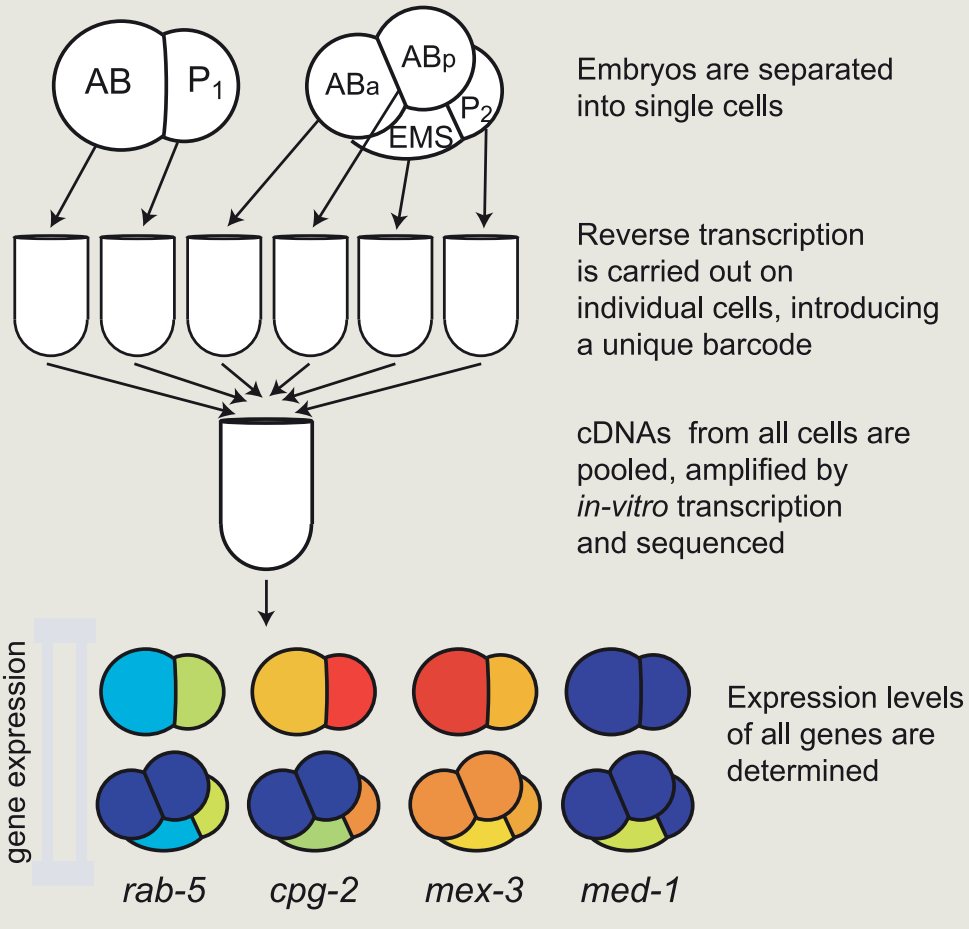


Figure 1. Hashimshony T, Wagner F, Sher N, and Yanai I (2012) CEL-Seq: Single cell RNA-Seq by multiplexed linear amplification. *Cell Reports, 2 (3): 666-673.*

EXPERIMENTAL DESIGN

3.

Sequencing protocols – Single-Read (SR) vs. Paired-End (PE), insert size, and read length

Deciding which sequencing protocol to choose is influenced by several factors:

The repetitive nature of the genome:
Human and mouse genomes are comprised of ~20% repetitive sequences. Consequently, in order to uniquely score a read mapping to a repetitive region, the read must be longer than the repetitive region or border the neighboring non-repetitive sequence. Thus, longer or PE reads facilitate accurate identification of a repetitive region's genetic location.

Differentially spliced variants:
When assessing gene expression levels in RNA-Seq, it is often important to identify differential expression levels of various transcripts of the same gene. Reads that map to an exon shared by more than one transcript pose a challenge to transcript-of-origin assessment. PE reads may solve this problem if one end of the sequenced fragment maps to an exon that is unique to one of the transcripts.

Genetic distance of the sequenced sample from the reference genome:
If the sequenced samples are genetically distant from the reference genome, then it is imperative to select a read length that can compensate for these inherently mismatched reads.

Identifying structural variations:
Structural variations in the genome, such as long insertions or deletions, inversions, and translocations, are best ascertained with PE reads.

***De novo* assembly:**
De novo assembly remains a notoriously challenging undertaking that often results in a genome consisting of thousands of contigs. Longer PE reads and sequencing multiple libraries of different insert lengths are two ways to improve *de novo* assembly.

Number of samples for sequencing

Resequencing:
If a sample's reference genome is genetically distant, then sequencing the strain in its baseline state (before mutagenesis, without the phenotypic change, etc.) will aid in data interpretation, including distinguishing variations due to evolutionary distance from those that cause the phenotypic trait of interest.

RNA-Seq:
It is highly recommended to sequence biological replicates in order to account for biological noise and improve statistical analyses.

ChIP-Seq:
A ChIP-Seq experiment should include the IP DNA and a control (input DNA or mock ChIP). Input DNA is DNA that has been purified, cross-linked, and fragmented under the same conditions as the IP DNA, whereas "mock" ChIP reactions are performed using a control antibody that reacts with an irrelevant, non-nuclear antigen (IgG control). By comparing IP DNA sequences to those of control DNA, one can differentiate between peaks that are significantly enriched due to immunoprecipitation versus those that have received higher coverage due to a sensitivity to fragmentation or other DNA-specific traits.

Sequence coverage

In reading applications, coverage corresponds to the number of reads that cover each base in the genome on average. Coverage can be calculated as:

$$\textbf{average coverage} = \frac{\textit{read length} \cdot \textit{number of mapped reads}}{\textit{genome size}}$$

Note that only the number of mapped reads should be included in the above calculation. The recommended coverage for identifying genomic variants is 30X or more, while *de novo* assembly requires a much higher coverage. The ideal coverage in any given project depends on the purpose and design of the experiment. For example, when re-sequencing a population containing a variety of heterogenic genomes, the coverage must be higher for the robust detection of rare variants.

Due to unequal read coverage in counting applications, such as RNA-Seq, there is no one formula for selecting the appropriate coverage for each project. In RNA-Seq, for instance, more highly expressed transcripts will receive higher coverage while lowly expressed transcripts will receive less coverage.

In these cases, it is recommended to evaluate transcriptomic complexity by beginning with a pilot experiment of just a few samples in order to assess what the ideal coverage for each individual application could be.

An example of an analysis that can help assess whether enough reads have been sequenced is a "saturation report" (Figure 2). In this "jack-knifing" method, the expression levels are determined using all of the reads. The expression levels are then compared to those recalculated using only a fraction of the reads. Examining the expression levels at each cut of the data is useful for identifying the point at which expression level remains unchanged despite additional data. As expected, additional data is helpful in resolving expression levels of lowly expressed genes. After determining the number of reads required per sample, the samples are divided into lanes according to the number of sequenced reads per lane, which is a fixed amount.

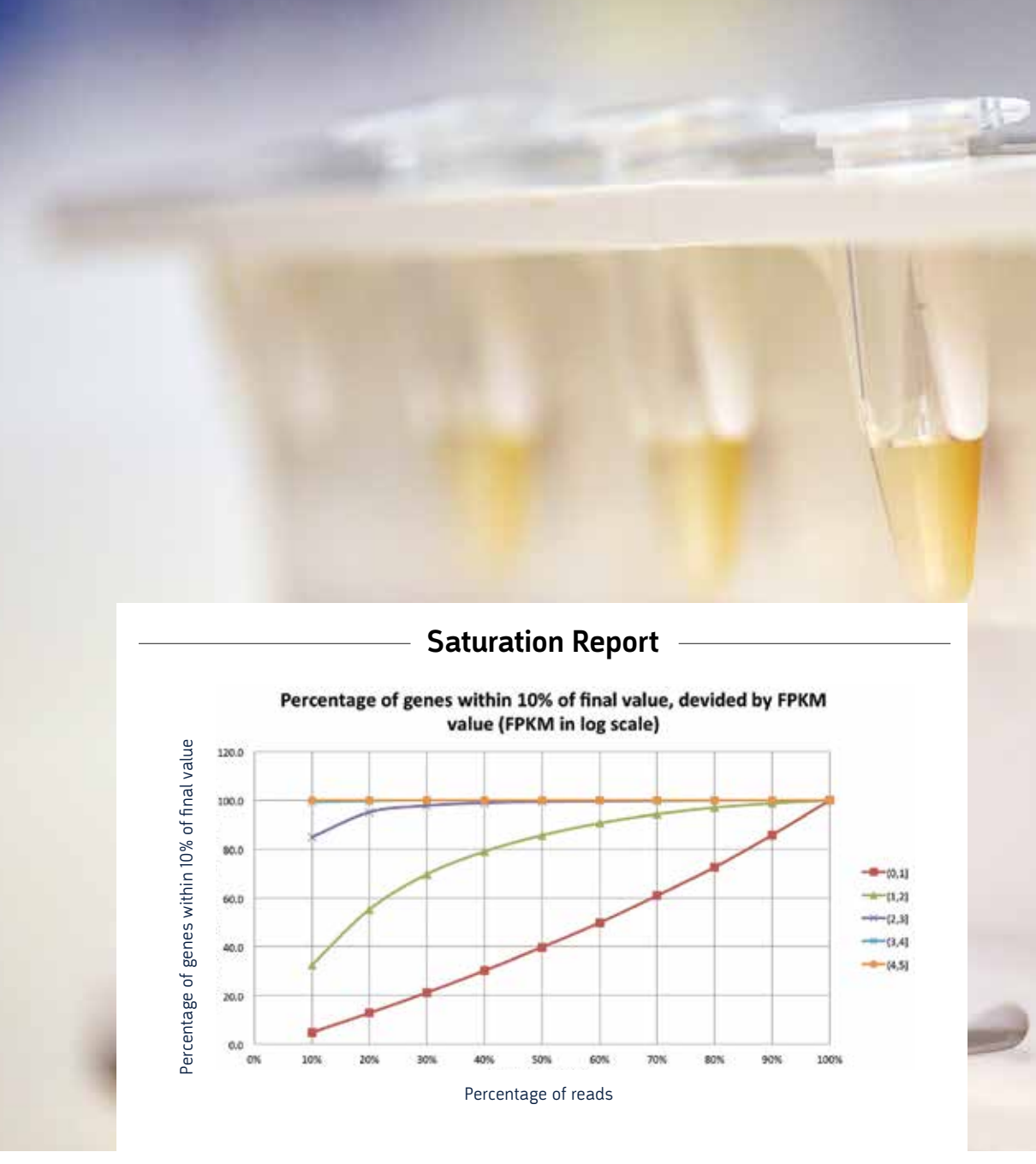


Figure 2. Each series is a set of genes that differ in their final expression values using the complete dataset (in this case, 32 million reads). Highly expressed genes are saturated with as little as 10% of the reads, whereas lowly expressed genes require a higher amount of reads. Very lowly expressed genes remain unsaturated even with the complete dataset. Figure and caption adapted from: "An introduction to high-throughput sequencing experiments: design and bioinformatics analysis," by R. Normand and I. Yanai, 2013, Deep Sequencing Data Analysis, *Methods in Molecular Biology*, 1038, p. 1-26.

TGC WORKFLOW 4.

Consultation meeting

We want to give you the best service possible! So, before submitting any samples for sequencing please contact us to set a meeting to discuss your project. Each consultation meeting is attended by a sequencing specialist and a bioinformatician.

Submission of samples for sequencing

After the meeting, samples can be submitted along with an approved sample submission form. The samples will then be processed for sequencing.

Sample preparation

Sample preparation is done either by the TGC team or by the researcher using any Illumina compatible protocol, in coordination with the TGC team.

Sequencing

Sequencing is conducted on our HiSeq2500 or MiSeq instruments, according to the project's requirements. Both sequencers are based on Illumina's sequencing by synthesis (SBS) technology. In most cases, libraries are compatible with both the HiSeq and the MiSeq.

Bioinformatic analysis

Data collection, processing, and analysis of the sequenced reads are achieved by use of a variety of software according to the required applications. These include the full range of data collection, processing, and analysis modules to streamline collection and analysis of data with minimal user intervention.

Report and concluding meeting

The researcher receives all of the raw data, initial analysis, and a detailed report describing the quality statistics and analyses conducted.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

What to expect when sequencing at the TGC

We at the TGC want your project to succeed! In order to provide the best service possible, we believe that it is important to coordinate expectations between the researcher and the TGC from the very beginning. Then, in order to achieve optimal results, we are in communication with researchers throughout the sequencing process—from the consultation meeting to sample preparation and sequencing, all the way through bioinformatic analysis.

What kind of service will you receive from the TGC?

Consultation meeting:
During the consultation meeting the researcher presents the biological question and together we plan the sequencing experiment. We explain the high-throughput sequencing technique and what to expect from the bioinformatic analysis. After the meeting, a sample submission form and an analysis questionnaire are filled out by the researcher in order to describe the sequencing and analysis specifications.

Concluding meeting:
At the concluding meeting, an overview of the final analysis results are presented to the researcher including data on the quality of the run and the bioinformatics pipeline that was used (software, parameters, etc.). The researcher learns how to use a genome viewer (such as IGV) in order to view and work with the results.

What will you receive from the bioinformatic analysis?

Report:
The concluding report includes a detailed explanation of the analysis pipeline, such as which tools, parameters, and statistics were applied at each step. In addition, the report contains a summary of the quality and technical details of the sequencing run.

Excel tables:
A summary table for each analysis application. For example, variants (for resequencing and exome analyses), differential gene expression (for RNA-Seq), peaks (for ChIP-Seq), and more. For further details, please refer to the "Bioinformatics" section.

Raw data:
All raw reads that were sequenced are given to the researcher. These files can be used for additional analyses.

IGV browser-compatible files:
With these files the researcher will be able to view the data at his/her leisure using the IGV genome viewer.

Additional application-specific results as discussed at the consultation meeting.

SAMPLE PREPARATION

Sample preparation is the process by which an initial sample, often genomic DNA or total RNA, is processed to become a library ready for sequencing.

5.

Consultation meeting

Sample preparation

Sequencing

Bioinformatic analysis

Report and concluding meeting

Preparation of genomic DNA samples begins with random shearing of the DNA, resulting in blunt-end fragments. These blunt ends are then adenylated in preparation for adaptor ligation.

Adaptors contain unique indexes to individually tag each sample for identification after sequencing. Size-specific magnetic beads are used for fragment size selection and then adaptor-bound fragments are enriched via PCR amplification. Enrichment of adaptor-bound fragments eliminates nonspecific ligation products and brings each sample to a working concentration that can be quantified for library normalization and sequencing.

If the starting material is total RNA, then samples undergo poly-A selection or ribosomal depletion in order to select for mRNA. The mRNA is fragmented, reverse transcribed to cDNA, and then undergoes a similar process to that of the DNA sample preparations.

The adaptors that were ligated to fragments during sample preparation hybridize to the flow cell on which they are sequenced. These adaptors contain a unique 6-8bp sequence, known as an "index" or "barcode, " essentially "tagging" each individual sample and making it possible to sequence multiple samples together as a pool. Because index sequences are unique, individual samples are then identified according to their assigned index during bioinformatic analyses.

There are many kits for preparing libraries to be sequenced by Illumina, such as the Illumina TruSeq and Nextera kits, NEB's NEBNext kits and more. Additionally, some labs prepare sequencing libraries using their own protocols.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Sample Requirements

Consultation meeting

Sample preparation

Sequencing

Bioinformatic analysis

Report and concluding meeting

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

DNA and RNA samples:
Just as Next-Generation Sequencing technology is constantly being improved and upgraded, so are library preparation kits and requirements.

The TGC always keeps up with the newest technology and applications, so please don't hesitate to contact us if:

- Your samples do not quite meet the requirements listed in the table below
- You have very little sample input
- You are interested in an application that is not listed

The table below is subject to change as protocols are continually upgraded.

Sample quality requirements:
Genomic DNA should be intact. If your sample is degraded please contact us to coordinate suitable sample prep.
RNA integrity should be confirmed using the Agilent Bioanalyzer/ TapeStation/similar instrument, or by running the sample on an agarose gel.

Sample purity requirements:
OD260/280 = 1.8-2.2
OD260/230 ≥ 2.0

Please note that the user is responsible for the sample's quality.

Sample Preparation Protocols and Requirements

LIBRARY PREPARATION TYPE	INPUT MATERIAL	MIN AMOUNT	VOLUME
TruSeq Nano DNA (standard DNA sample prep)	gDNA	200 ng	Up to 50 µl
Nextera XT (small genomes and amplicons*)	gDNA of small genomes/ amplicons	5 ng	Up to 10 µl
Exome Sequencing	gDNA	1 µg**	Up to 50 µl
ChIP-Seq	ChIP DNA	20 ng	Up to 50 µl
TruSeq RNA (standard RNA sample prep)	Total RNA/ mRNA	1 µg	Up to 50 µl
ScriptSeq complete (Stranded RNA sample prep with ribosomal depletion)	Total RNA	1 µg	Up to 50 µl
CEL-Seq	aRNA	Please contact us	Please contact us
SMARTer and SMART-Seq sample preps	Total RNA/ mRNA	Please contact us	Please contact us

**For amplicon sequencing please contact us. Amplicons will be accepted only following DNA purification.*
***For low input protocol, please contact us.*

User Prepared Library Requirements

Please submit at least 10µl of library (2-10 nM) suspended in 10 mM Tris-Cl, pH 8.5. For orders consisting of two or more libraries to be sequenced together in a single lane (HiSeq or all MiSeq orders), please submit samples as a pool.

If possible please submit Bioanalyzer or TapeStation traces.

We use standard illumina sequencing primers, and read a single index of 6bp. Please inform us if your libraries:

- Require different sequencing primers (please check with us for the primer's availability or include custom primers when submitting your libraries)
- Require longer index reads or dual index reads
- Have any special characteristics (low diversity, unbalanced, poly A, etc.)

TGC-prepared sequencing libraries undergo standardized quality assessment and calibration to ensure optimal cluster generation and densities. User-prepared libraries cannot be guaranteed optimal cluster densities, though many come very close. However, some deviate significantly from ideal cluster density, thereby compromising the number of reads. In these situations, the user will still be billed for sequencing.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

ILLUMINA SEQUENCING TECHNOLOGY

6.

Sequencing Overview

Illumina's innovative and flexible sequencing system enables a broad array of applications in genomics, transcriptomics, and epigenomics.

Libraries are prepared from genomic DNA or RNA, then immobilized on the surface of a flow cell designed to present the DNA in a manner that facilitates access to enzymes, while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labeled nucleotides.

Solid-phase amplification creates 1,000 identical copies of each single template molecule in close proximity with total cluster densities on the order of 10^6 clusters/mm^2.

Consultation meeting

↓

Sample preparation

↓

Sequencing

↓

Bioinformatic analysis

↓

Report and concluding meeting

Sequencing by Synthesis

Sequencing by synthesis (SBS) technology uses four fluorescently labeled nucleotides to simultaneously sequence the tens of millions of clusters on the flow cell surface.

During each sequencing cycle, a single fluorescently labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. The nucleotide label serves as a terminator for polymerization.

After each dNTP incorporation the fluorescent label is imaged to identify the base and later enzymatically cleaved to allow incorporation of the next nucleotide. Since all four reversible terminator-bound dNTPs (A, C, T, and G) are present as single separate molecules, natural competition minimizes incorporation bias. Base calls are made directly from signal intensity measurements during each cycle, thereby reducing raw error rates.

The end result is highly accurate base-by-base sequencing that eliminates sequence-context specific errors, enabling robust base calling across the genome.

BIOINFORMATIC ANALYSIS

7.

Bioinformatic analyses extract the results from raw sequencing data. The researcher receives various tables to summarize and visualize these results. Those tables are thus the basis for downstream research.

The analysis pipeline

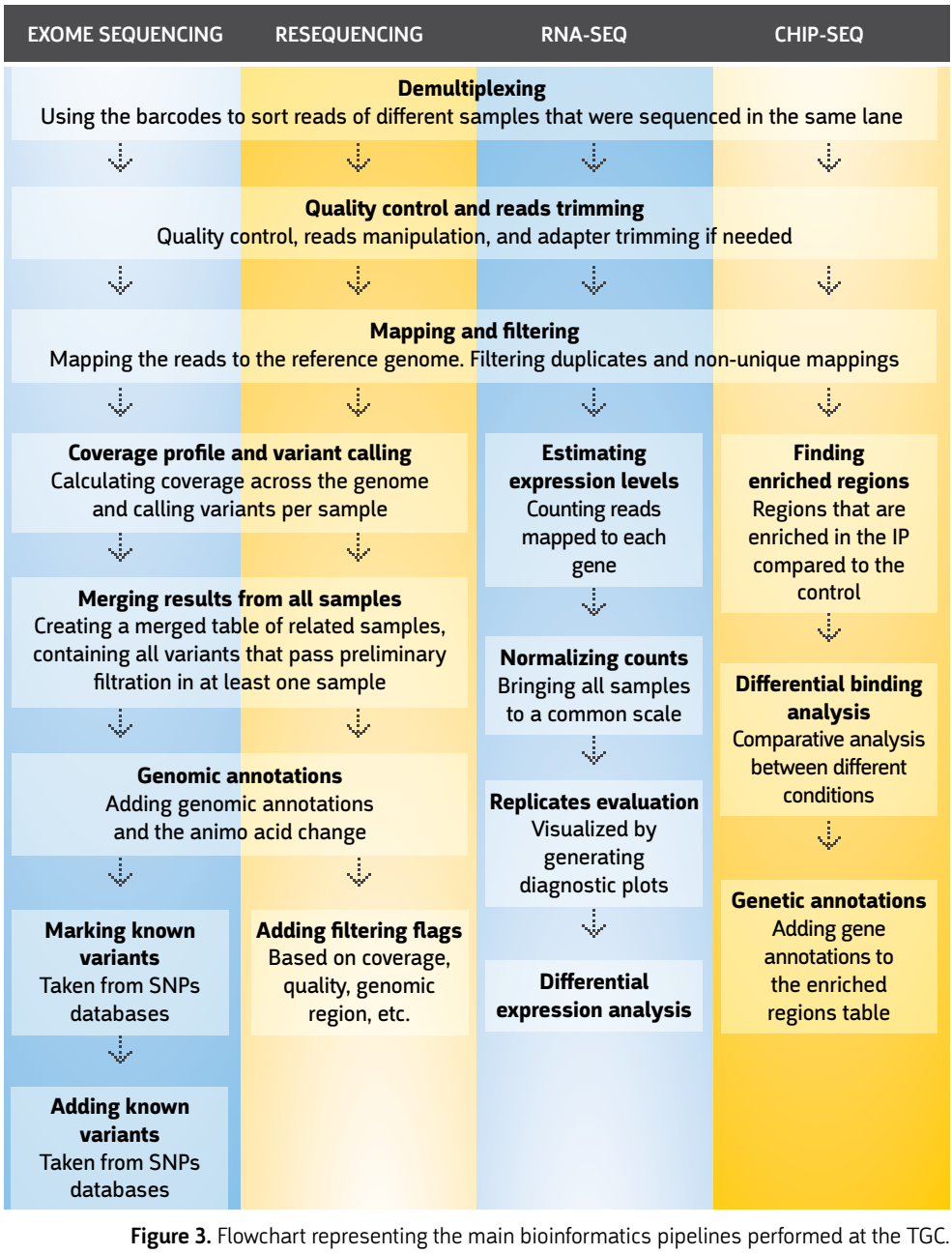
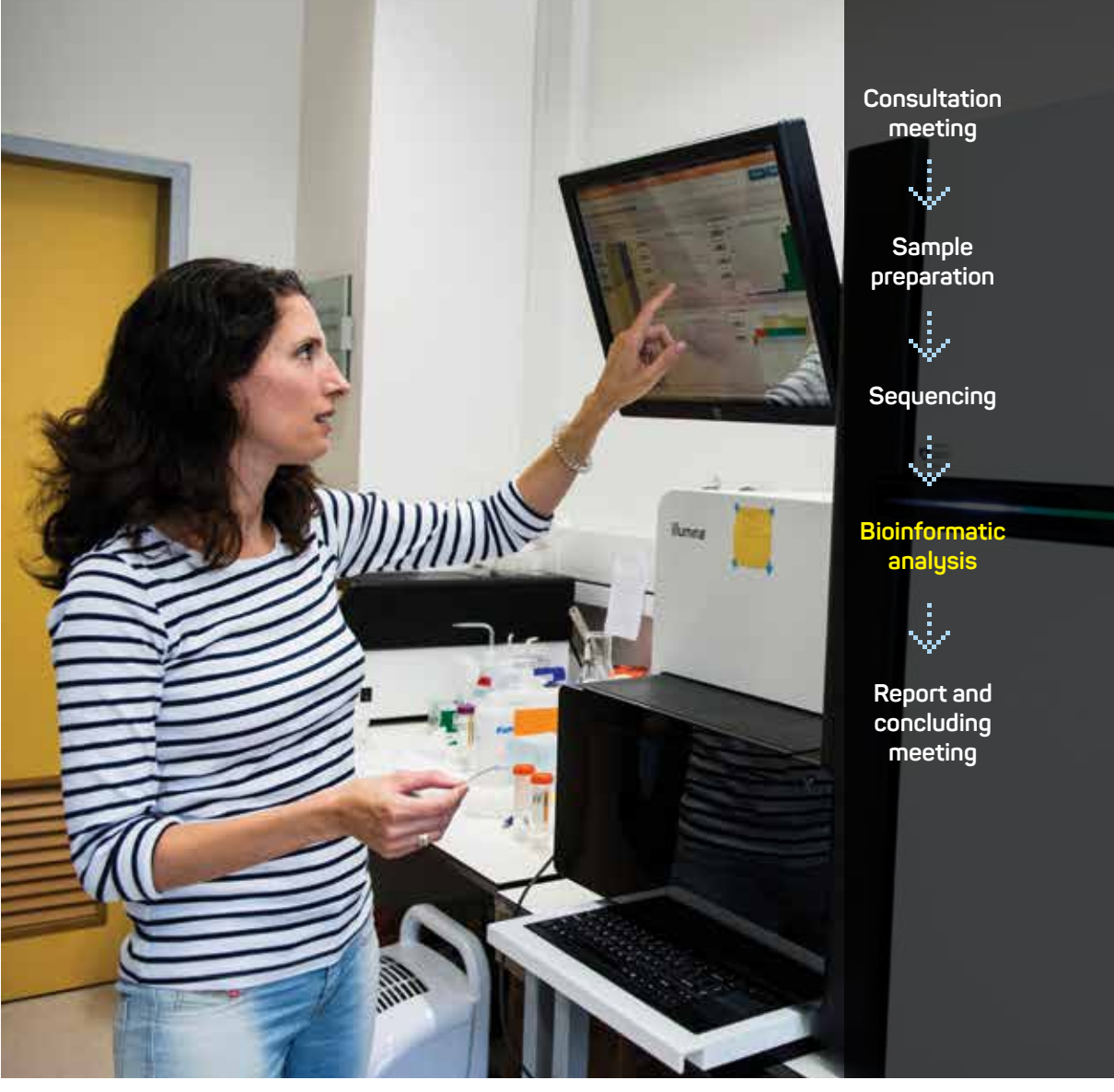


Figure 3. Flowchart representing the main bioinformatics pipelines performed at the TGC.



What is included in the final bioinformatics report?

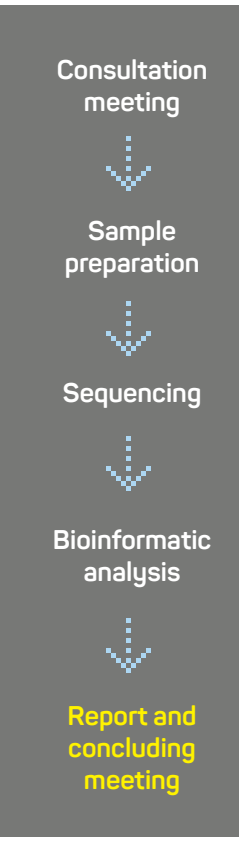
For each project we provide the researcher with a final report including a detailed explanation of the analysis pipeline, such as which tools and parameters were used and the statistics of each step. In addition, the report contains a summary of the quality and technical details of the sequencing run.

The final results files of the analysis are comprised of tables summarizing each analysis application, the raw data for the researcher's future use, IGV browser-compatible files, and any additional application-specific results as discussed at the consultation meeting.

Results tables for each type of analysis

For your convenience, technical details are provided in the form of a table detailing all of the pertinent traits of each sample as seen in the following analysis segments.

Resequencing analysis



The researcher is provided with a "variants table" detailing the differences between each sample's sequence and the reference genome, such as SNPs, insertions, and deletions.

The specific information provided for each project includes:

Chromosome/Position:
The location of each variant.

Reference:
The base or sequence of the reference genome at the position of the variant.

Allele(s):
The detected base or sequence(s).

Variant Type:
Indicates whether the detected variant is a SNP, insertion or deletion.

Phred Quality score for allele call:
The quality score of the allele call, representing the probability that the detected variant exists at the site of interest.

Genomic annotations:
Available genomic annotations depend on the organism sequenced.

Filtering flags:
The filtering criteria are custom made for each analysis according to the researcher's specifications (not present in figure 4).

For each sample:

Genotype:
The genotype call.

Coverage:
The total number of reads that were mapped at the position of interest.

Strand count:
For each strand, the number of reads supporting the reference sequence and the number of reads supporting the non-reference sequence at that position.

Allele frequency:
Calculated as the non-reference reads count divided by total number of high quality reads at the position of interest.

Genotype Quality:
The confidence level of the genotype assignment for the variant.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Variants Characteristics - Resequencing Analysis

a.

CHROMOSOME	POSITION	REFERENCE	ALLELE/S	VARIANT TYPE	PHRED Q-SCORE FOR ALLELE CALL
Chr1	533181	A	G	SNP	65.19
Chr1	1033896	C	T	SNP	405.47

b.

GENOTYPE	ALLELE GENOTYPE	HIGH QUALITY READS Ref, Alt	FREQUENCY	COVERAGE	GENOTYPE QUALITY	STRAND COUNTS (Ref fw, Ref rev, Alt fw, Alt rev)	MAPPING QUALITY	MAPPING QUALITY U Test
0,1	A,G	100,40	0,4	140	99	45,55,18,22	60	-0,34
1,1	T,T	0,1360	1	1,360	99	0,0,690,670	40	-

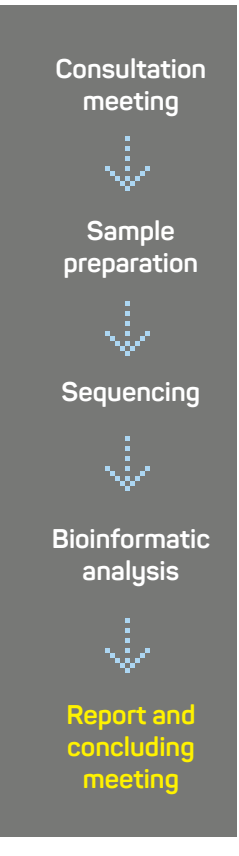
c.

EFFECT	TYPE	CODON CHANGE	AMINO ACID CHANGE	GENE	BIOTYPE
NON_SYNONYMOUS_CODING	MISSENSE	Aag/Gag	K914E	AT1G02530	protein_coding
SYNONYMOUS_CODING	SILENT	agG/agA	R374	AT1G04010	protein_coding

Figure 4. Example of a resequencing analysis variants table.

a. General information about the detected allele(s).
b. Sample-specific information. The columns pictured above appear in the table for each of the analyzed samples.
c. Genomic annotations of the variant position in the genome. The specific annotations provided in this table differ according to the organism's database.

Exome analysis



Researchers electing to conduct exome analyses receive a merged table of variants for each set of related samples (such as family members analyzed for a specific disease causing mutation).

This merged table consists of data as previously described in the Resequencing section, with additional fields specific to exome analyses (see list below).

Reference Amino Acid:
The amino acid(s) derived from the healthy allele(s).

Allele/s Amino Acid:
The amino acid derived from the detected allele.

Region type:
Exonic/intronic/UTR, etc.

Annotation:
Whether the codon change is synonymous or nonsynonymous.

Gene:
The name of the gene at the position of interest.

Sift score:
Numerical representation predicting the effect of amino acid change on the protein function.

SNPdb / 1,000Genomes:
The ID of known variants at the position of interest.

Filtering flags:
The expected heritage type (such as autosomal recessive, etc.) and information on the relations between samples is used to mark the relevant variants according to genotype. Additional filtering criteria are added, such as quality, coverage, region type, etc.

In addition, **exon coverage** is calculated and reported in a separate table with combined coverage statistics.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Variants Characteristics - Exome Analysis

a.

CHROMOSOME	POSITION	REFERENCE	ALLELE/S	PHRED Q-SCORE FOR ALLELE CALL
1	2494330	G	A	60

b.

GENOTYPE SAMPLE 1	ALLELE GENOTYPE SAMPLE 1	HIGH QUALITY READS Ref, Alt SAMPLE 1	FREQUENCY SAMPLE 1	COVERAGE SAMPLE 1	GENOTYPE QUALITY SAMPLE 1	STRAND COUNTS (Ref fw, Ref rev, Alt fw, Alt rev) SAMPLE 1	MAPPING QUALITY SAMPLE 1	MAPPING QUALITY U Test SAMPLE 1
0,1	G,A	120,121	0,5	241	99	50,70,60,61	60	0,67

GENOTYPE SAMPLE 2	ALLELE GENOTYPE SAMPLE 2	HIGH QUALITY READS Ref, Alt SAMPLE 2	FREQUENCY SAMPLE 2	COVERAGE SAMPLE 2	GENOTYPE QUALITY SAMPLE 2	STRAND COUNTS (Ref fw, Ref rev, Alt fw, Alt rev) SAMPLE 2	MAPPING QUALITY SAMPLE 2	MAPPING QUALITY U Test SAMPLE 2
1,1	A,A	0,190	1	190	87	0,0,100,90	60	-

c.

REFERENCE AMINO ACID	ALLELE/S AMINO ACID	REGION TYPE	ANNOTATION	GENE	TRANSCRIPT/S	SIFT SCORE	SNPdb	1000Genomes	GENOTYPE SAMPLE 1 - Aa, SAMPLE 2 - aa
M	T	exonic	nonsynonymous	TMEM52	NM_178545	0.52 SNP	rs2860257:A->G	.: A->G	TRUE

Figure 5. Example of an exome analysis variants table.

a. General information about the detected allele(s).
b. Sample-specific information. The columns pictured above appear in the table for each of the analyzed samples.
c. Genomic annotations of the variant position in the genome. IDs of known variants in published databases and filtering flags. Filtering flags are project-specific; the main filtering criteria are selected based on information provided by the researcher.

RNA-Seq analysis

RNA-Seq analysis results include normalized read counts for each gene, as well as differential gene expression analysis results between different biological conditions.

The information provided to the researcher in the differential expression results table includes:

Gene ID and position:
For each gene.

baseMean:
The average normalized expression level across all analyzed samples.

For each requested comparison between conditions:

Fold Change:
Log_2 of the fold change between the expression levels of compared conditions.

P-value:
The uncorrected p-value.

Padj:
The adjusted p-value.

Flag:
An indicator showing whether or not the gene passed a certain minimum expression level.

Normalized Counts Columns:
The detected normalized counts of all replicates of the compared conditions.

Significantly DE (Differentially Expressed):
An indicator representing whether or not the gene passed a certain threshold of significance according to the adjusted p-value.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Differential Gene Expression Analysis

a.

GENE ID	GENE NAME	GENE POS	baseMEAN
ENSG00000086015	MAST2	1:45786987-46036124	0.921356674
ENSG00000211769	TRBJ2-5	7:142797119-142797166	430.4921647
ENSG00000101442	ACTR5	20:38748442-38772520	504.2159982

b.

LOG2FOLDCHANGE	FOLDCHANGE	PVALUE	PADJ	FLAG
-0.113760015	0.924176292	0.798135014	NA	Low_Counts
-1.494064536	0.355010959	0.000400676	0.01861908	Tested
-0.084893613	0.w	0.845255377	0.926555164	Tested

c.

NORMALIZED COUNTS CONDITION A	NORMALIZED COUNTS CONDITION B	SIGNIFICANTLY DE
1.17;0	0;2.31;1.12	no
226.08;140.55	674.53;642.51;468.79	yes
483.78;484.6	770.8;407.54;374.36	no

Figure 6. Example of differential gene expression analysis results from RNA-Seq data.

a. General information on each gene. The subjects displayed in this table depend on the organism of interest. The information in figures b. and c. are provided for each requested comparison between conditions.
b. Statistical results of the differential expression analysis.
c. The normalized expression levels of all samples relevant to each comparison are shown in columns grouped according to replicate sets. An additional indicator, Significantly DE (Differentially expressed), is provided. This value is based on a minimal threshold on the adjusted p-values.

ChIP-Seq analysis

ChIP-Seq is an application used to analyze protein-DNA interactions. It combines chromatin immunoprecipitation (ChIP) with high-throughput DNA sequencing to identify histone modification locations or the binding sites of DNA-associated proteins, such as transcription factors.

The type of experiment and specifications requested by the researcher determine which tools and pipelines to use for analysis, generating several types of final results files.

All results tables include the coordinates of the detected enriched regions, statistical conclusions, and genomic annotations.

Consultation meeting

Sample preparation

Sequencing

Bioinformatic analysis

Report and concluding meeting

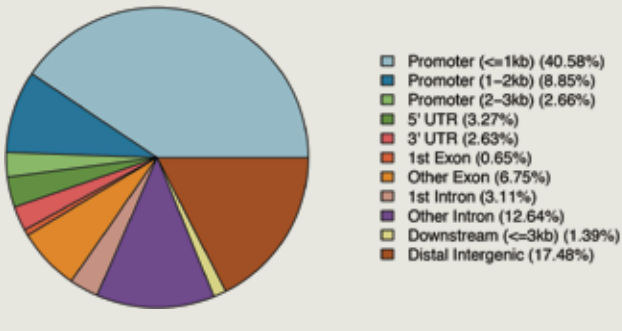


Figure 7. Pie chart of genomic annotations of the transcription factor (TF) of interest, examining the location distribution within the genome.

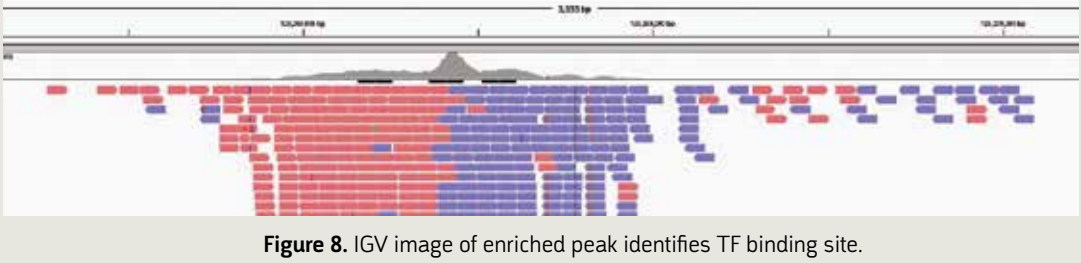


Figure 8. IGV image of enriched peak identifies TF binding site.

We're here to answer your questions:

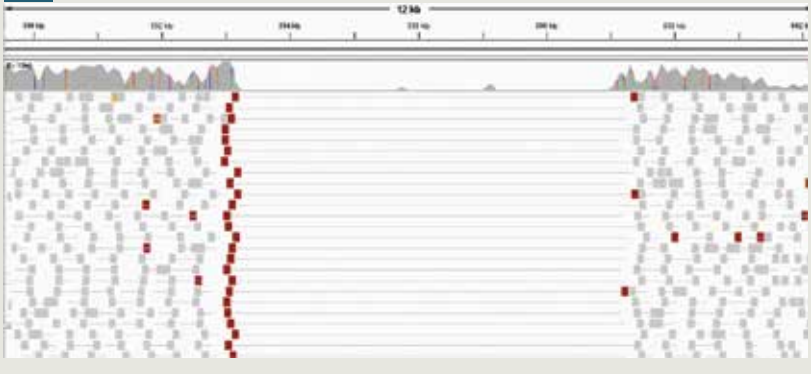
Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

IGV files

IGV is a free genome viewer that allows the user to visualize the data. IGV can be used to see mapping results, coverage, and allele consensus.

The TGC provides the researcher with IGV-compatible files that can be used in conjunction with the TGC-provided results tables in order to identify and visualize a region of interest and conduct downstream analyses.

a.



b.

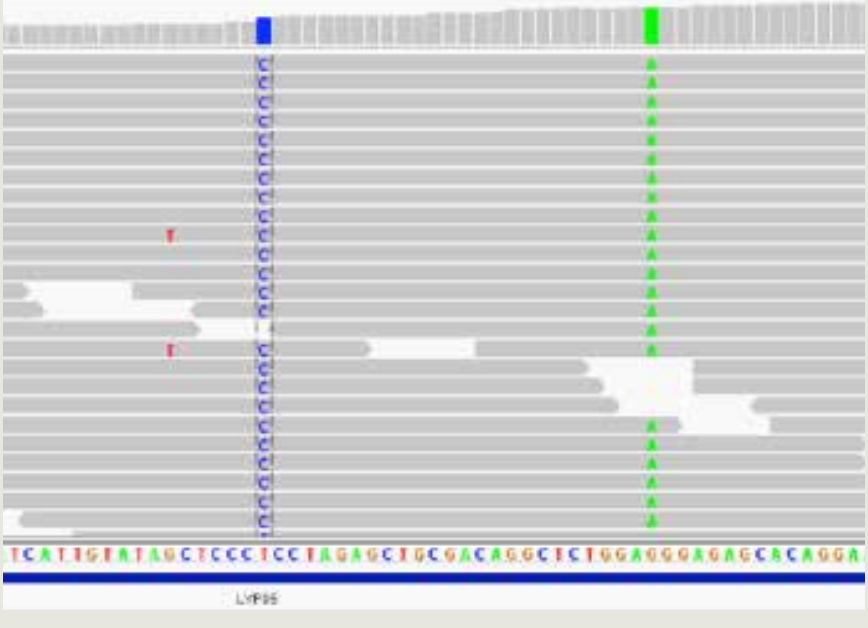


Figure 9. Examples of mapping results shown using IGV.

Understanding the complexity of sequencing analysis

Consultation meeting

Sample preparation

Sequencing

Bioinformatic analysis

Report and concluding meeting

As described in this chapter, the details of each analysis differ according to each project's unique specifications. In order to optimize analysis results, it is crucial to determine the pipeline best suited for each project individually. Additionally, each step of the pipeline must be adjusted to the project's unique specifications in order to obtain accurate results. This can be achieved by following the general pipeline presented in this chapter and conducting quality control measurements after each step.

Quality control measurements should be conducted immediately after each step of the analysis. Two such quality control assessments are: mapping statistics and coverage.

Mapping statistics help determine the number of reads that were not mapped, uniquely mapped, and multi-mapped. High percentages of unmapped and multi-mapped reads may be indicative of problematic sequencing libraries. It is highly recommended to look at the mappings in a genome viewer.

Coverage profile assessments include examining the profile visually, the percentage of the genome with sufficient coverage, and average coverage. For exome projects, an additional parameter to be checked is the exon coverage. Some phenomena can be detected easier visually (see Figure 10). In the example provided in Figure 10, two bacterial samples were sequenced and mapped to the same reference genome. The mapping statistics of both revealed that 96-98% of the reads were unmapped. Visualizing the results on a genome viewer reveals the differences between the two samples as shown below.

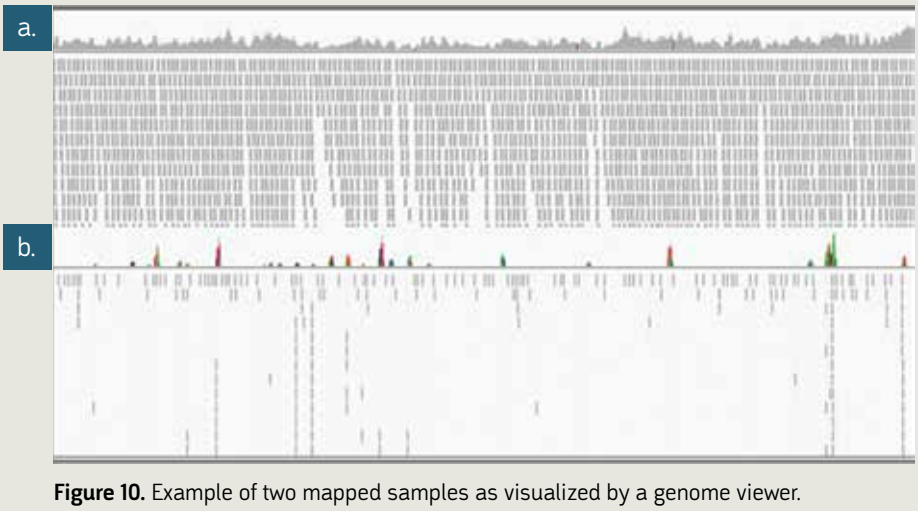


Figure 10. Example of two mapped samples as visualized by a genome viewer.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Sample a shows a continuous and high coverage, while **sample b** shows discontinuous coverage with many variants. How can this be explained?

Sample a. This sample initially appears to have particularly high coverage, however the continuous coverage and lack of variants indicate that only 2-3% of the reads map to the expected strain. Quality assessment of the coverage profile thus characterizes a sample that is 97-98% contaminated.

Sample b. At first glance this sample appears to have very low coverage, potentially indicating problematic libraries. Further examination, however, reveals that the sample's low and segmented coverage is due to a high incidence of variants. Taken together, the coverage profile thus characterizes a sample that is evolutionarily distant from its reference genome.

Tuning the parameters of each step of the analysis makes it possible to control the balance between sensitivity and specificity. For example, if one mismatch per 50bp read is allowed in the mapping step, the rate of incorrect mappings will be reduced, but 2-base indels or areas in the genome that have more than one variant per 50bp will not be detected. Therefore, coverage in these regions will be low or zero due to inefficient mapping.

When comparing gene expression between two samples one can choose to statistically test only genes that have a minimum amount of reads mapped in at least one sample.

Choosing a high threshold may cause elimination of interesting genes, but choosing a low threshold may include genes whose differential gene expression is not significant. For example, if the ratio between samples is defined as 1:5 as opposed to 1,000:5,000, then gene expression of a given gene will be five-fold higher in the latter analysis.

EQUIPMENT

Just as improvements in Next-Generation Sequencing are constantly being made, instrument performance is perpetually improving.

The TGC always keeps up with the newest technology and applications, so please feel free to contact us for up-to-date information.

8.

HiSeq 2500

We have two Illumina HiSeq 2500 instruments. The HiSeq sequencing system uses Illumina's proven reversible terminator-based sequencing by synthesis technology, delivering ultra-high-throughput sequencing and fast data generation. It can be operated in single or dual flow cell mode, allowing applications requiring different read lengths to run simultaneously.

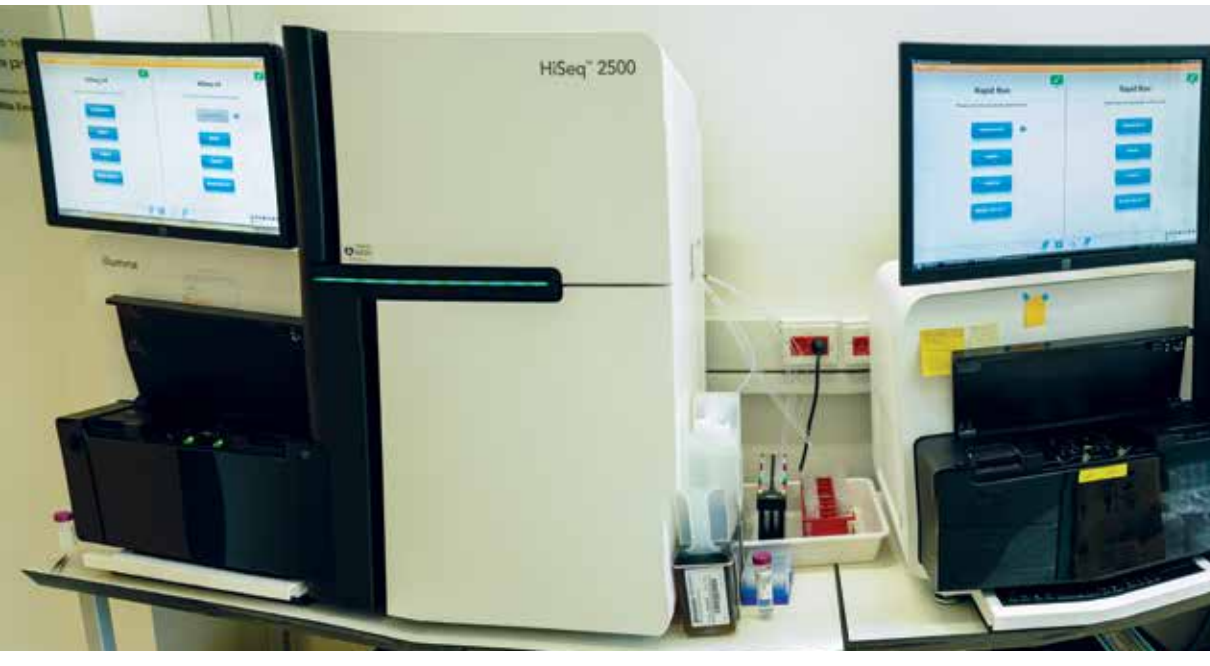
New HiSeq V4 reagents allow for more reads and more data in less time. The HiSeq 2500 features two run modes: High Throughput Mode and Rapid Run Mode.

HiSeq Performance Specifications

READ LENGTH	HIGH OUTPUT	RAPID RUN MODE
Max Read Length	2X125	2X250
Max run length	6 days	60 hr
Reads	Up to 2 billion single reads or 4 billion paired-end reads	Up to 300 million single reads or 600 million paired-end reads

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il



DID YOU KNOW?

The HiSeq 2500 generates up to ~280 million reads per lane and ~560 million paired-end reads per lane.

A whole human genome can be sequenced at 40X coverage using only two lanes of a 125bp paired-end run in less than a week.

MiSeq

The Illumina MiSeq is a desktop sequencer used for sequencing small genomes, assemblies, amplicons, and other applications that require longer read length and fewer reads.

- Enables longer reads, up to 300bp PE.
- Produces up to 25 million SR reads and 50 million PE reads.

DID YOU KNOW?

The MiSeq generates up to ~25 million reads and ~50 million paired-end reads per run.

We're here to answer your questions:

Technion Genome Center
Tel: 972-4-829-5168
tgc@tx.technion.ac.il

Covaris E-220

The Covaris E-220 is a multi-sample DNA shearing system. It is a Focused Ultrasonicator designed for shearing of genomic DNA and chromatin.

The Adaptive Focused Acoustics (AFA) process employs focused bursts of ultrasonic acoustic energy at a frequency of 15 to 30 times higher than that of a sonicator.

The AFA technology allows:

- Extraordinary reproducibility due to tight parameter control.
- Higher yield and better quality due to effects of applying focused acoustics.

The E-220 allows multi-sample processing, treating up to 96 samples per use, each with its own unique parameter specifications.

Agilent 2200 TapeStation

Using the same basic principles as gel electrophoresis, the Agilent 2200 TapeStation allows for fast and simple quality assessment of RNA and DNA samples. Only 1-2µl are required from each sample, and results are obtained within ~1 minute per sample.

DNA screen working concentrations:
Standard DNA: 0.1-50 ng/µl,
High sensitivity DNA: 10-1000 pg/µl

RNA screen working concentration:
Standard RNA: 25-500 ng/µl,
High sensitivity RNA: 500-10,000 pg/µl

The TapeStation software automatically calculates the RNA integrity number equivalent (RINe) for total RNA samples, thus providing an objective measurement of RNA quality and degradation.

The TapeStation is also used to analyze prepared libraries prior to sequencing. The results of running two prepared DNA libraries are shown in Figure 11 (on facing page).

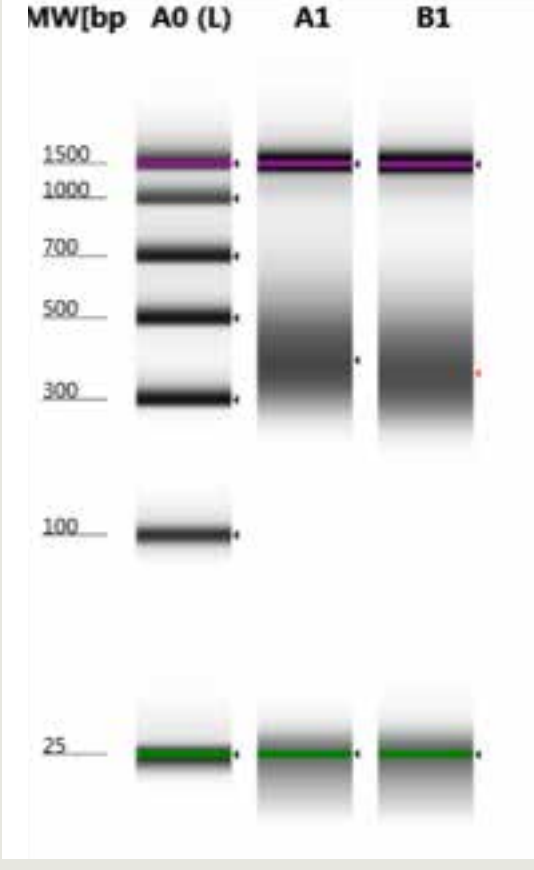
Figure 11. The TapeStation software provides two visual representations of the data from each run:

a. The genetic material as it appears on the electrophoretic gel.

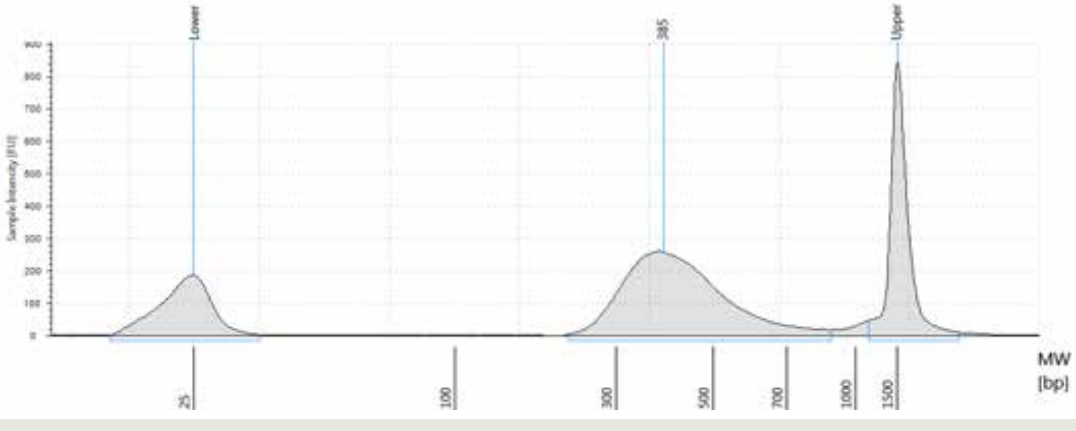
b. A graphical representation depicting the fluorescent intensity and molecular size of each sample.

The TapeStation software identifies and measures significant peaks and allows the user to define regions of interest for calculating average sizes of each region.

a.



b.



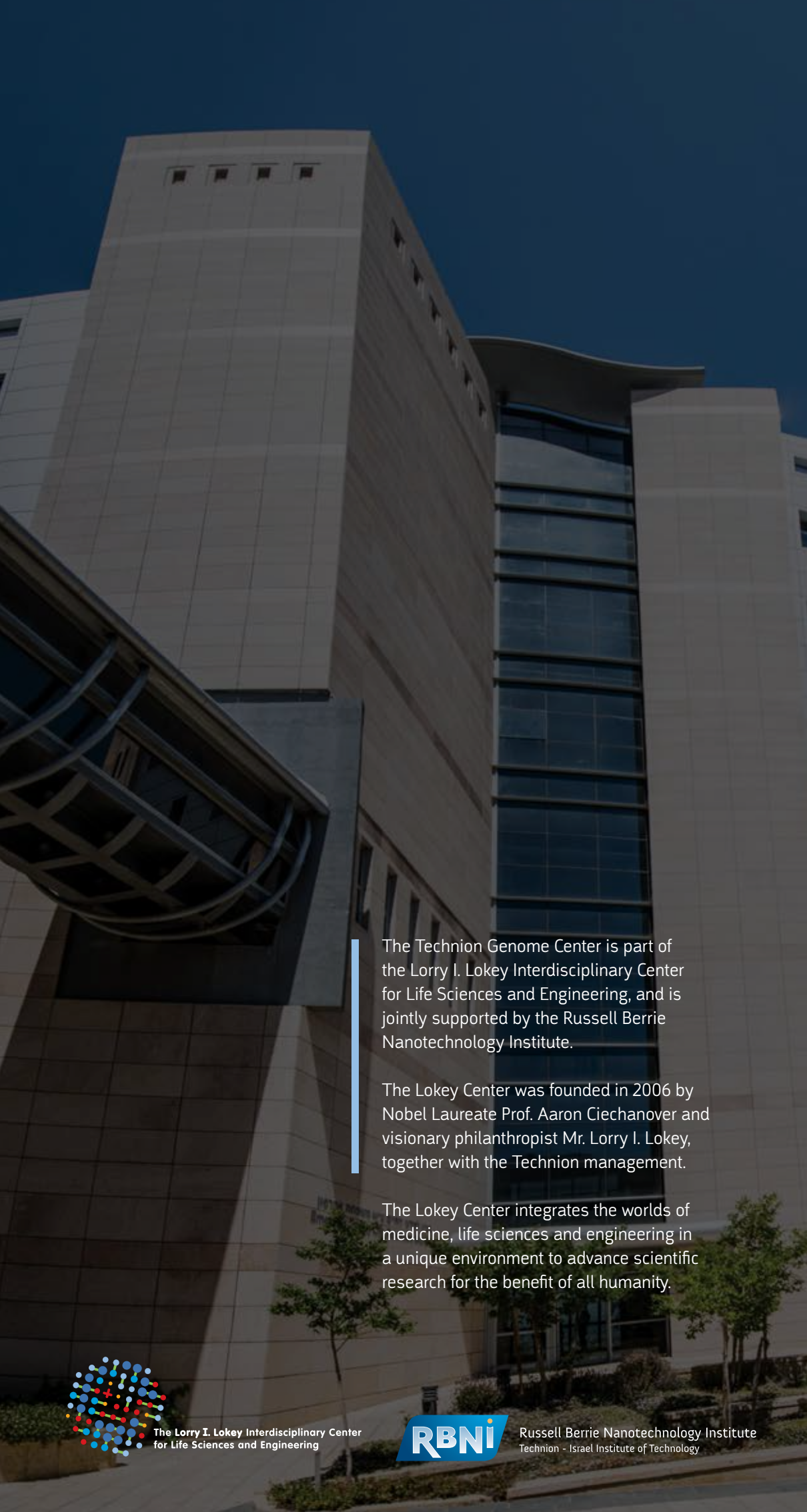
MW [BP]	CONC. [PG/µL]	MOLARITY [PMOL/L]	OBSERVATIONS
25	382	23100	Lower Marker
385	1020	4010	
1,500	580	586	Upper Marker

Agilent Bravo Automation System

For high-throughput library preparations the TGC uses the Agilent Bravo automated liquid handling system.

With the Bravo platform we can prepare up to 96 DNA or RNA samples in a single, automated run, significantly reducing library preparation time.

The Bravo system is reliable and precise, producing high quality libraries in a fraction of the time.



The Technion Genome Center is part of the Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering, and is jointly supported by the Russell Berrie Nanotechnology Institute.

The Lokey Center was founded in 2006 by Nobel Laureate Prof. Aaron Ciechanover and visionary philanthropist Mr. Lorry I. Lokey, together with the Technion management.

The Lokey Center integrates the worlds of medicine, life sciences and engineering in a unique environment to advance scientific research for the benefit of all humanity.

The **Lorry I. Lokey** Interdisciplinary Center for Life Sciences and Engineering

RBNI Russell Berrie Nanotechnology Institute
Technion - Israel Institute of Technology

TECHNION GENOME CENTER (TGC)
Emerson Building for Life Science,
Technion
Haifa 32000 Israel
Tel: 972-4-829-5168
Fax: 972-4-8295131

tgc@tx.technion.ac.il
http://tgc.net.technion.ac.il

TECHNION
Israel Institute of Technology