Novogene

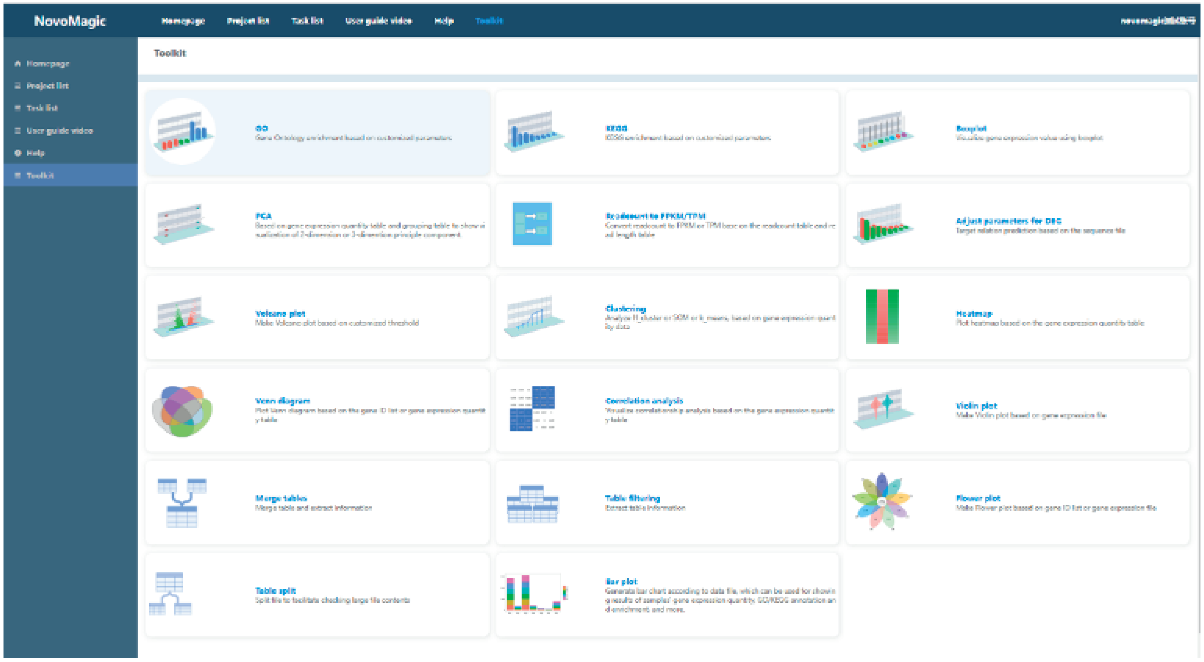
Do Transcriptome Analysis With Just a Click

Novogene

CONTENTS

Introduction

Transcriptome sequencing, also referred to as RNA sequencing (RNA-seq) or massively parallel RNA-seq, is a powerful tool that offers deep insights into gene expression and how changes in gene expression drive biological processes and impact health.[1,2] However, RNA-seq experiments generate large amounts of data that must be carefully managed and analyzed. RNA-seq data analysis is a complex pipeline requiring different types of data processing, quality control (QC), and statistical analysis at each stage. Naturally, those necessary analyses require specialized software tools. Here, we briefly discuss several of the analytical steps involved, and how you can use the tools of Novogene's NovoMagic platform to accelerate your RNA-seq data analysis. [3]



GO
Gene Ontology enrichment based on customized parameters

KEGG
KEGG enrichment based on customized parameters

Boxplot
Visualize gene expression value using boxplot

PCA
Based on gene expression quantity table and grouping table to show visualization of 2-dimension or 3-dimension principle component

Readcount to FPKM/TPM
Convert readcount to FPKM or TPM base on the readcount table and real length table

Adjust parameters for DEG
Target relation prediction based on the sequence file

Volcano plot
Make Volcano plot based on customized threshold

Clustering
Analyze H_cluster or SOM or k_means, based on gene expression quantity data

Heatmap
Plot heatmap based on the gene expression quantity table

Venn diagram
Plot Venn diagram based on the gene ID list or gene expression quantity table

Correlation analysis
Visualize correlationship analysis based on the gene expression quantity table

Violin plot
Make Violin plot based on gene expression file

Merge tables
Merge table and extract information

Table filtering
Extract table information

Flower plot
Make Flower plot based on gene ID list or gene expression file

Table split
Split file to facilitate checking large file contents

Bar plot
Generate bar chart according to data file, which can be used for showing results of samples' gene expression quantity, GO/KEGG annotation and enrichment, and more.

Fig 1. Novogene's NovoMagic platform

NovoMagic is a free cloud-based platform specially designed and developed by Novogene for our customers. Through re-analysis functions, NovoMagic enables you to rename your samples, re-analyze gene expression, adjust the parameters of differentially expressed genes, and perform gene function analysis. It features 17 toolkits to help you easily customize your RNA sequencing data for publishing in just a few clicks.

1. Visual display and normalization of transcript expression

After RNA-seq data has been generated and the raw sequence reads have been subjected to initial QC processing and mapped to a reference transcriptome or genome, the next stage in the analysis pipeline is quantitative analysis or expression analysis.[1] The goal is to quantify gene expression.

1.1 Read counts and expression density

The simplest approach to quantifying gene expression is creating read counts—counting the number of cleaned reads that map (align) to each gene. [1-3] Read counts can be obtained using the Gene Filtering tool in the NovoMagic platform.[3] Although read counts themselves do not directly correspond to gene expression data that can be used experimentally, read counts can provide insight into the nature and quality of the RNA-seq data that has been obtained

Read counts can be analyzed for coverage and depth. Coverage refers to the number of times a reference base in a reference genome has been "covered" by sequenced fragments. Depth refers to the percentage of bases of a reference genome that has been "covered" by short reads at a certain depth.[3]

Applying some basic descriptive statistics to a sample's read count also reveals characteristics of the data such as density and distribution. Consistent density of read counts indicates that the data is reliable and is more normally distributed. A histogram of counts per gene, preferably log2-transformed counts for improved visibility, shows the distribution of counts per gene. Quality read count data is expected to be consisted with a negative binomial distribution. A density plot, a smoothed version of the histogram, is helpful for visualizing the read count distributions of multiple samples at the same time. Box plots are useful for visualizing the gene count distribution per sample.[4,5]

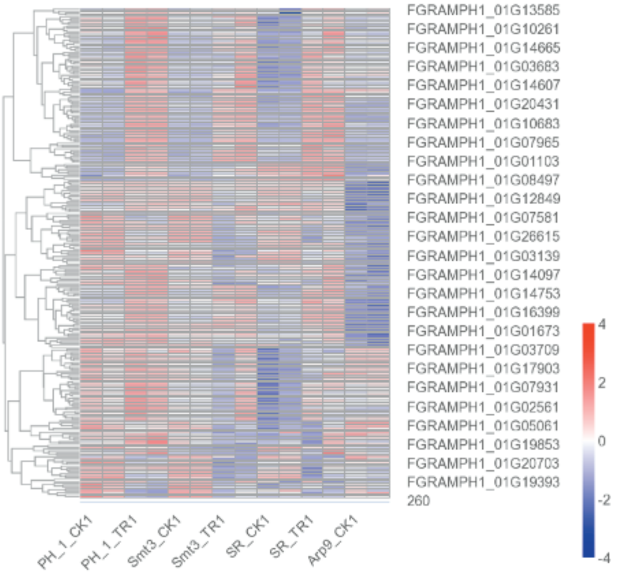


Fig 2*. Heatmap of test data

1.2 Transcript abundance measures and normalization

Identifying relevant changes in gene expression, increases and decreases, is an important function of RNA-seq analysis. Downstream analyses and comparisons of those gene expression changes can then investigate whether the expression changes are of biological importance. While RNA-seq fragment counts can be used as a measure of relative transcript abundance, read counts alone cannot be used to accurately evaluate or compare gene expression within or between samples.[1,2,6,7]

Normalization methods scale raw read counts to make expression levels more comparable within or between samples. Normalization attempts to account for several factors, including sequencing depth (necessary to compare expression between samples), gene length (necessary to compare expression of different genes within a sample), and RNA composition (important for comparing expression between samples and for differential expression analysis). [7]

The three most common RNA-seq normalization methods for comparing gene expression within a sample are RPKM (reads per kilobase of exon per million reads mapped), FPKM (fragments per kilobase of exon per million fragments mapped), and TPM (transcripts per kilobase million). [1-3,6,7] RPKM is used for single-end RNA-seq experiments, while the analogous FPKM measure is used for paired-end RNA-seq experiments. RPKM/FPKM and TPM measures are not recommended for comparisons between samples or for differential gene expression (DGE) analysis,[1,6,7] although TPM may be used for comparing gene expression levels between samples of the same sample group.7 The NovoMagic platform allows users to convert from read count to FPKM or TPM.[3]

However, RNA-seq normalization methods must be used judiciously, and there is currently no consensus regarding the most appropriate method for cross-sample comparisons. [6] To compare gene expression between samples, differential gene expression (DGE) and its accompanying statistical methods are recommended.[1]The NovoMagic platform support to normalize RNA-seq data and analyze DGE by DESeq2 or edgeR. With the adaptable nature of the NovoMagic platform, users have options to use the normalization method best suited to their individual needs.

1.3 Correlation analysis

Once normalization for within-sample transcript comparisons is complete, correlation analysis should be performed as an additional quality control (QC) step.[1] Correlation analysis is often performed using clustering analysis methods to compare the transcript profiles of all samples in the data set.[1] The expectation is that biological replicates should cluster together because they should have similar transcript profiles. Likewise, biological replicates should be less similar to samples representing different biological conditions because different biological conditions would be expected to have different transcript profiles. The purpose is to determine intra- and inter-group variability for samples and to identify outliers that were not excluded during earlier QC steps. [3]

Two common methods for correlation analysis are hierarchical clustering and principal component analysis (PCA). Hierarchical clustering results are often represented with a heat map, which assigns data values to a range of colors. When arranged as a matrix in which rows and columns are sorted by similarity, a heat map shows patterns of high and low correlation that indicate which samples are most similar to one another. [1,8] Principal component analysis (PCA) identifies the features (dimensions) that explain most of the variance between transcript profiles.[1,3] NovoMagic platform users have the option of representing correlation data with either heat maps or PCA plots between samples or group.

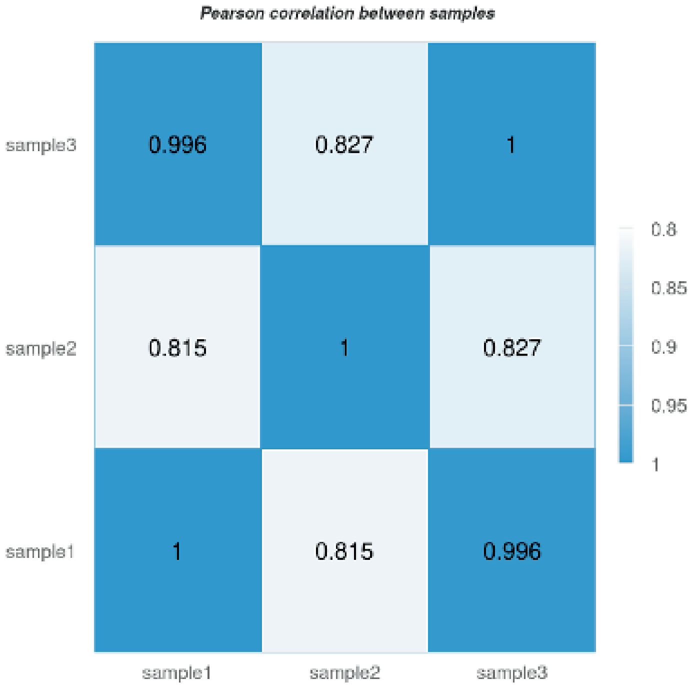


Fig 3*. Correlation between samples

2. Differential gene expression analysis

DGE analysis, also referred to as differential expression (DE), is a fundamental goal of RNA-seq analysis. It is also important to undertake DGE analysis carefully because subsequent analyses (see, for example, section 3 below) usually rely on DGE results. Because DGE analysis aims to identify transcripts with altered expression levels in one experimental group compared to the expression levels in another group by comparing the deviation between a group mean and a global mean. [1,3,10] For example, DGE analysis can be used to identify transcripts that are enriched (have a higher expression) or depleted (have a lower expression) in one tissue versus another tissue. It is also often used to identify tissue-specific expression —genes that are expressed in one tissue or a set of tissues but not expressed in other tissues.[1]

Differential expression is measured using fold change (FC), which is usually reported on a log2 scale. A positive FC value indicates increased expression, and a negative FC value indicates decreased expression.[10] Of course, a change in expression alone is not necessarily meaningful; it also needs to be statistically significant. Because DGE analysis of an RNA-seq experiment involves separately testing many genes for differential expression, a multiple hypothesis testing burden must be accounted for—that is, there is a greater likelihood of false positive results.

Therefore, an adjusted p-value should be used when evaluating statistical significance. [1,10]NovoMagic users have two options for DGE analysis: DESeq2 for data sets that include replicates and edgeR for data sets without replicates. Both are statistically designed for comparisons among samples and use adjusted p-values.

There are also a variety of ways to display DGE results graphically. Several popular options are volcano plot, Venn diagrams, and bar plot.1 Volcano plots are useful for comparing up-or down-regulation of transcript expression of all genes.[1,3]Because Venn diagrams show the intersections between or among two or more groups, they can be used to visualize the amounts of shared and not shared expression among gene sets.[1,3] Histograms are useful for showing the distribution of transcript expression levels under various conditions.[1,3]All these mentioned plots could visualization by NovoMagic.

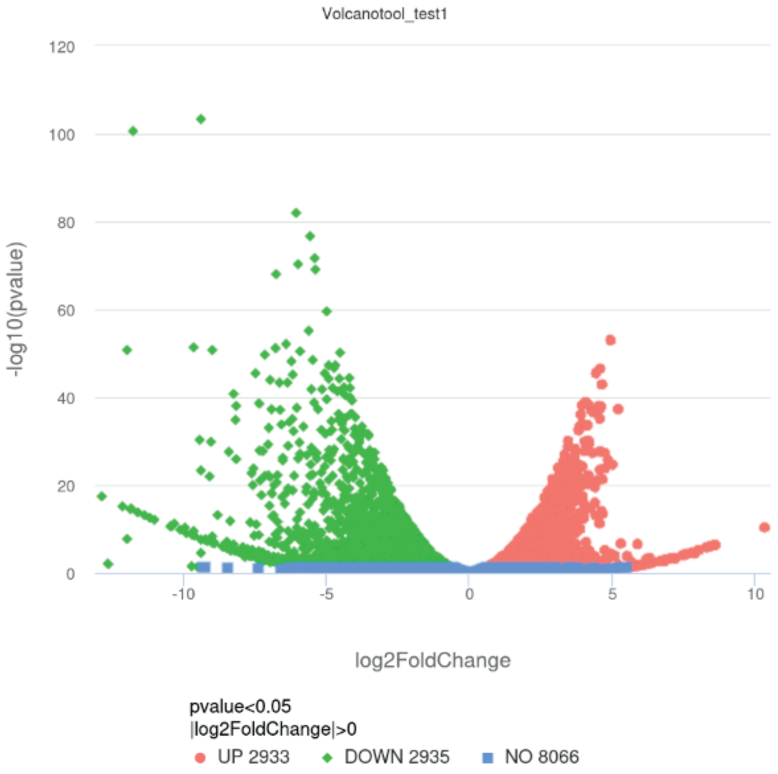


Fig 4*. Volcano plot of test data

3. Gene set enrichment analysis

After DGE analysis, the next step in RNA-seq analysis is often gene set enrichment analysis, assigning biological meaning to gene sets of interest by accurately placing RNA-seq transcript expression data into the appropriate biological context(s).[1] This approach can confirm existing relationships among gene expression and biological conditions, such as gene expression during a disease state. It can also identify potentially novel relationships, thus driving hypothesis formation and discovery. Enrichment analysis investigates whether a gene set is enriched for genes that share a biological feature, such as a signaling pathway, a protein function, or a biological response.[1] Here, we briefly discuss three common methods: gene ontology enrichment, KEGG pathway enrichment, and network path analysis.[1,11,13,17]

3.1 Gene ontology (GO) enrichment

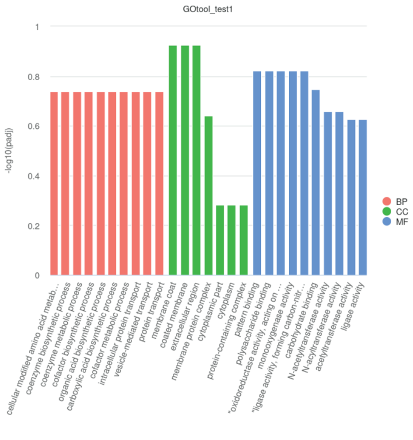


Fig 5*. GO plot of test data

An ontology provides standardized language for categorizing and cataloging knowledge within a field of study.[11] The Gene Ontology (GO), maintained by the Gene Ontology Consortium, consists of (1) a hierarchical ontology of concepts (GO terms) and (2) the GO annotation. The hierarchical ontology specifies biological elements molecular function, biological process, cellular component. The GO annotation lists annotated genes that are linked to GO ontology terms.[11] Thus, GO enrichment analysis of a gene set is a tool for connecting differentially expressed genes that have been annotated to biological processes via the GO.

However, investigators should keep in mind that the GO and its annotations are not static: they are continuously changing and are subject to annotation bias (for example, Tomczak et al. noted as recently as 2018 that over 50% of annotations were for less than 20% of all human genes). Therefore, investigators should use prudence when interpreting GO enrichment analyses and should revisit earlier analyses using the most recent GO version.[11]

NovoMagic users can use the platform's clusterProfiler tool for GO enrichment analysis and construct GO enrichment scatter plots.3 GO enrichment analyses can also be visualized using a type of circos plot called a chord diagram. Genes and GO terms are arranged around the plot circumference; the genes are connected by lines (chords) to the GO terms for which they are enriched.

3.2 KEGG mapping

The Kyoto Encyclopedia of Genes and Genomes (KEGG) "is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism, and the ecosystem, from genomic and molecular-level information." [12] It represents biological systems as molecular wiring diagrams of networks that integrate genomic, chemical, reaction, relational, and health information.[12] KEGG mapping for RNA-seq experiments is therefore the process of mapping differentially expressed genes to KEGG molecular networks, and thereby generating a new set.[13] The best-known type of KEGG molecular network is the KEGG pathway map, which represents a molecular network by KEGG Orthology (KO) groups so that it can be generalized among organisms.

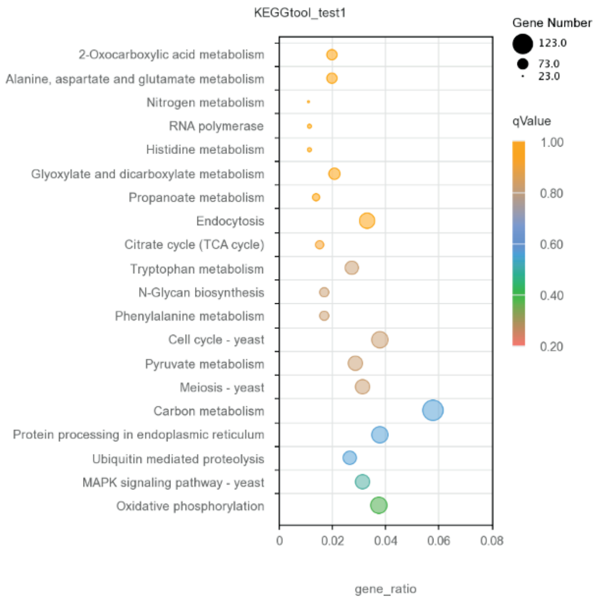


Fig 6*. KEGG plot of test data

Genes can also be mapped to KEGG BRITE hierarchies (which incorporate multiple types of relationships including genes and proteins, compounds and reactions, drugs, diseases, and organisms and viruses) and KEGG modules. [13,14,15] Thus, KEGG mapping is a method for identifying putative high-level functions of differentially expressed genes. The NovoMagic platform's clusterProfiler tool also performs KEGG mapping.

NovoMagic users may use the "all-in-one" module in the transcriptome analysis process, which requires just one set of parameters and grouping information, and can perform differential analysis and gene enrichment with a single click. Novogene customers can try out some of the guest mode features for free to get started with RNA sequencing data analysis. If you don't currently have an active project with Novogene, you can contact us via email below to get a demo account for NovoMagic.

✉ America: inquiry_us@novogene.com
✉ AMEA: marketing_AMEA@novogeneait.sg
✉ Europe: pm@novogene-europe.com

References

1. Cockrum C et al. A primer for generating and using transcriptome data and gene sets. Development. 2020;148;dev193854. doi:10.1242/dev. 193854

2. Griffith M et al. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLoS Comput Biol. 2015;11;e1004393. doi:10.1371/journal.pcbi.1004393

3. Introducing NovoMagic—Novogene's Online RNA-seq Bioinformatics Analysis Tool. Updated November 24, 2022. Accessed February 6, 2023. https://www.novogene.com/us-en/resources/onlineevent/introducing-novomagic-novogenes-online-rna-seq-bioinformatics-analysis-tool/#

4. Puthier D., adapted from Varet H, Auberta J, and van Helden J. RNA-Seq - differential expression using DESeq2. Github.io. Published December 10, 2016. Updated January 23, 2023. Accessed February 6, 2023. https://dputhier.github.io/ASG/practicals/rnaseq_diff_Snf2/rnaseq_diff_Snf2.html

5. Pine Biotech. Visualizing Gene Expression: Box plots, histograms, and density plots in R. https://www.youtube.com/watch?v=UXJFu0ddp1E Published October 20, 2020. Accessed February 6, 2023.

6. Zhao Y et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med. 2021;19;269. doi:10.1186/s12967-021-02936-w

7. Harvard Chan Bioinformatics Core. Introduction to DGE-Archived. Github.io. Accessed February 6, 2023. https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

8. JMP® Genomics Users Guide. Correlation Heat Map (Correlation and Principal Components). Accessed February 6, 2023.
https://www.jmp.com/support/downloads/JMPG101_documentation/Content/JMPGUserGuide/OT_G_EX_0006.htm

9. Nygaard V et al. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016;17;29-39. doi: 10.1093/biostatistics/kxv027

10. Comparing experimental conditions: differential expression analysis. Github.io. Accessed February 6, 2023. https://biocorecrg.github.io/CRG_Bioinformatics_for_Biologists/differential_gene_expression.html

11. Tomczak A et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci Rep. 2018;8;5115. doi: 10.1038/s41598-018-23395-2

12. KEGG Overview: 1. Genomes to Biological System. Kyoto Encyclopedia of Genes and Genomes. Updated November 1, 2022. Accessed February 6, 2023. https://www.genome.jp/kegg/kegg1a.html

13. KEGG Mapping. Kyoto Encyclopedia of Genes and Genomes. Updated June 10, 2019. Accessed February 6, 2023. https://www.genome.jp/kegg/kegg1b.html

14. KEGG Pathway Maps. Kyoto Encyclopedia of Genes and Genomes. Updated June 1, 2011. Accessed February 6, 2023. https://www.genome.jp/kegg/kegg3a.html

15. KEGG BRITE Database. Kyoto Encyclopedia of Genes and Genomes. Updated January 1, 2023. Accessed February 6, 2023. https://www.genome.jp/kegg/brite.html

16. Creixell P et al. Pathway and Network Analysis of Cancer Genomes. Nat Meth. 2015;12;615-21. doi: 10.1038/nmeth.3440

17. Weidmann C et al. Analysis of RNA-protein networks with RNP-MaP defines functional hubs on RNA. Nat Biotechnol. 2021;39;347-56. doi: 10.1038/s41587-020-0709-7

*：Sources from Novogene Analysis Report

Novogene is committed to be

Your Trusted Genomics Partner

- Trusted International Corporation
- Trusted Comprehensive Experience
- Trusted Expertise & Technology
- Trusted Service & Standardized Procedures
- Trusted Delivery Quality & Efficiency

Follow us on LinkedIn

Novogene Co., Ltd

www.novogene.com

Novogene Global | Novogene_Global | Novogene Global

✉ inquiry_us@novogene.com | marketing@novogene-europe.com | marketing_amea@novogeneait.sg

Information and specifications are subject to change at any time without notice. Please contact your Novogene representative.