Information Flow from Strand NGS 2.1 to GeneSpring 13.0

Technical Overview

Authors

Srikanthi Ramachandrula[1], Prateek Singh[1], Maria Kammerer[1], Carolina Livi[2], Vanessa Lordi[2] and Pramila Tata[1]

[1]Strand Life Sciences
Kirloskar Business Park
Bangalore, India

[2]Agilent Technologies, Inc.
5301 Stevens Creek Blvd,
Santa Clara, CA, 95051, USA

Introduction

Studying complex biological systems requires the application of various high-throughput analytical technologies. To gain comprehensive insights into any biological phenomenon, it is essential to integrate data from various studies carried across these multiple platforms, as was described by Rhodes *et al.*[1], "...integrative approaches are capable of simplifying complex cancer signatures into co-ordinately regulated modules, transforming one-dimensional cancer signatures into multi-dimensional interaction networks and extracting regulatory mechanisms encoded in cancer gene expression." With rapid advances in all fields of biology, be it genomics, transcriptomics, proteomics, or metabolomics; integrative multi-omics analysis is the sure way forward.

New in GeneSpring 13.0 is support for the import of Next Generation Sequencing (NGS) experiments from Strand NGS 2.1 (http://www.strand-ngs.com/) facilitating Integrated Biology (IB) analysis using data from microarray, mass spectrometry, and sequencing platforms. The ability to analyze NGS experiments along with existing experiments in GeneSpring is an all important utility enabling:

- Identification of common pathways or pathway entities that are of significance in gene expression, metabolite abundance, and sequencing experiments.
- Correlation Analysis between sequencing studies and microarray expression data or metabolite abundance measured using mass spectrometry.
- Visualization of sequencing and array data in the same project.
- Cross-platform investigation leading to exploratory analysis.

The first two utilities are carried out in the context of a multi-omics experiment (with the NGS experiment being one of the inputs to the multi-omics experiment), while the last two do not necessarily need creation of a multi-omics experiment. This Technical Note describes use cases for a new integrated analysis workflow using high-throughput microarray and next generation sequencing data made possible by bridging Strand NGS 2.1 and GeneSpring 13.0.

Agilent Technologies

Export of Sequencing Experiments from Strand NGS 2.1

To perform a multi-omics analysis in GeneSpring involving a sequencing experiment, the sequencing analysis should be done in Strand NGS 2.1. With the exception of alignment, all the remaining experiments can be exported out of Strand NGS and imported into GeneSpring. To export relevant objects from Strand NGS, right-click on the experiment name in the project navigator to expose the **Export Experiment for GeneSpring** option. This launches the export dialog (Figure 1) wherein the objects to be exported can be chosen. Space requirements for the process are computed and displayed so that the user can choose to proceed further or, if necessary, reduce the resulting export file size.

To be noted

- Export from Strand NGS is a selective export at experiment-level wherein the objects to be exported from a given experiment are judiciously chosen based on user discretion. Therefore, it is different from **Export Project** where one or more experiments from a given project are bundled in their entirety.
- The exported experiment is an .expt file, which is distinct and different from a project export .tar file.
- Alignment experiments cannot be exported.
- RNA-Seq and smallRNA-Seq experiments can be exported only after quantification.
- A DNA-Seq or ChIP-Seq experiment created without Gene and Transcript annotations cannot be exported as there is no technology associated with it.
- Complete all the data analysis required for multi-omic analysis in the NGS experiments in Strand NGS before choosing to export so that all the corresponding read lists, entity lists, and region lists can be exported to GeneSpring together as a single experiment.

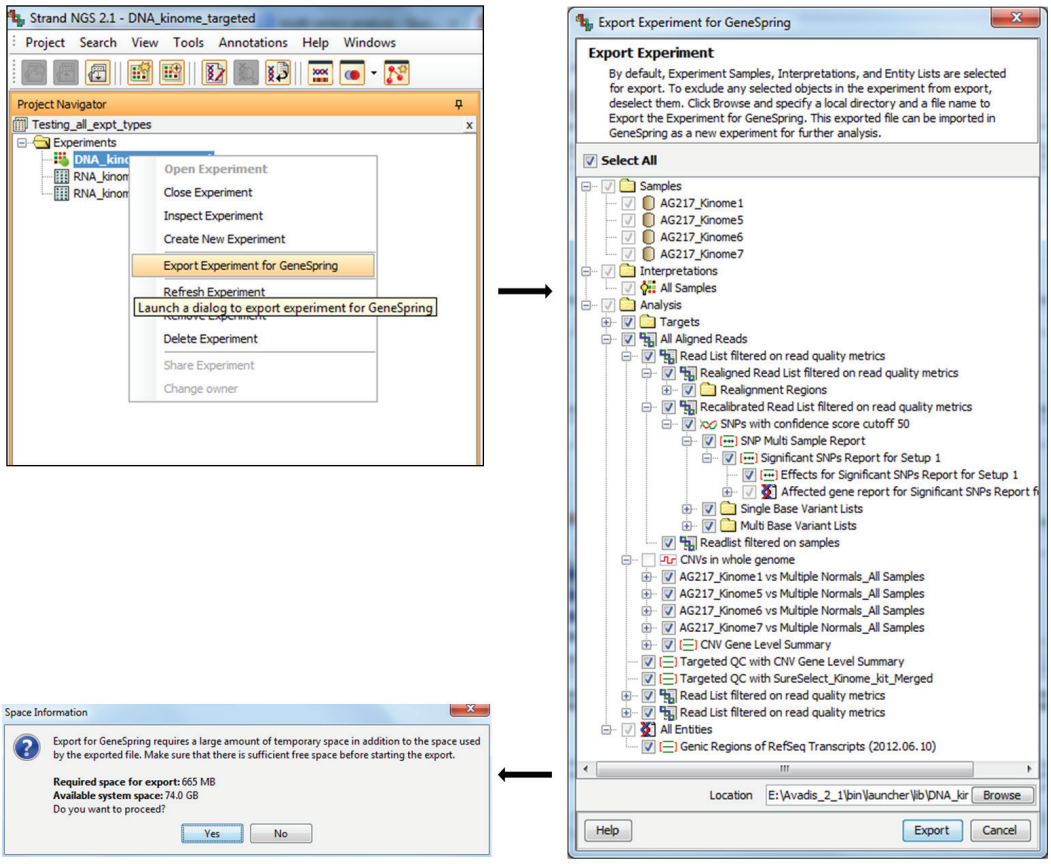

Figure 1. Export of sequencing experiment from Strand NGS 2.1.

Import of the NGS Experiment into GeneSpring

Clicking on the **Project Menu** in an open project in GeneSpring exposes the **Import Strand NGS Experiment** option. This launches a dialog permitting selected experiment files to be imported (Figure 2). Space requirements for the process are then computed and displayed so that the user can take necessary action.

To be noted

- The supported file format for experiment import into GeneSpring is .expt, exported from Strand NGS.
- Since the import is that of a single experiment (unlike project import), the relevant project in GeneSpring must be open for experiment import to proceed.
- Use the same build and annotations across GeneSpring and Strand NGS for seamless analysis; for example, if a data set were aligned to hg19 and quantified using gene and transcript model from UCSC (dated 2014.09.09) in Strand NGS, the same Reference and Gene model should be downloaded for use in GeneSpring.

Requirements – Space and Time

- The space requirement for export is roughly five times the size of the exported file. This is due to the creation of temporary objects during the export process. The temporary objects are cleared and the space is released at the completion of the experiment export.
- The space requirement for import is roughly two times the size of the file to be imported. Similar to the export process, the extra space used during import is released upon successful completion of experiment import.
- The time required for export is dependent on the number of objects to be exported, their composition, and the extent of overlap in their contents. Thus, two nonoverlapping read lists would take a longer time to export than if the read lists were overlapping. A given number of reads spread over several read lists would take longer to export as opposed to a single read list with the same number of reads.
- Time required is slightly higher when re-aligned and recalibrated read lists are exported.
- Bench-marking: The export of 14 GB (~150 million reads) of data takes approximately 2 hours, and importing the same takes approximately 40 minutes on a 64-bit Windows 7 machine with 8 GB RAM.

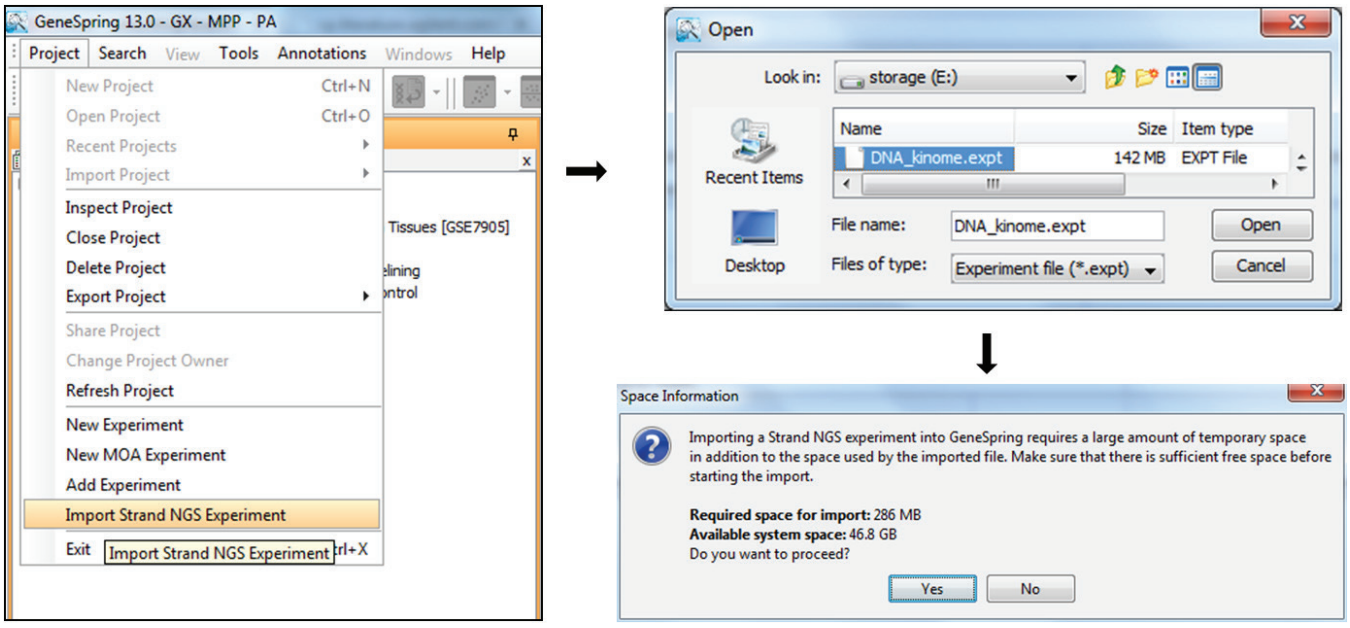

Figure 2. Importing the Strand NGS experiment into GeneSpring 13.

Objects to be Exported

Strand NGS supports the analysis of DNA-Seq, RNA-Seq, Methyl-Seq, SmallRNA-Seq, and ChIP-Seq. Apart from samples, interpretations, read lists, region lists, and entity lists, each of these experiments has a unique set of objects, which are generated depending upon the operation performed. Thus, CNV Pair Node, Continuous Data Node, Copy Number Node, and CNV Regions are created only if a copy number analysis is done in a DNA-Seq experiment. Similarly, Quantification Node, Novel Detection Report, Gene Fusion Node, and Gene Fusion Report will be created only upon performing quantification in an RNA-Seq experiment. Quantification Report and Active Regions are generated after quantification in a SmallRNA-Seq experiment.

Upon exporting a Strand NGS experiment, samples, interpretation, and entity list are exported by default, and the export of region lists is at the user discretion. Table 1 lists the objects created in different NGS experiment types and their export status.

Maintenance of Object Hierarchy

In Strand NGS, as in GeneSpring, a parent-child hierarchy relationship is maintained for the objects, that is, the read lists, entity lists, and region lists. This hierarchy is visible in the project navigator. If all the objects from the original experiment in Strand NGS 2.1 are exported, the object hierarchy is mirrored in the imported experiment in GeneSpring. When objects are selectively chosen for export, the imported experiment creates a hierarchy that is closest to the original experiment. This is illustrated in Figure 3, showing the export dialog of Strand NGS 2.1, and a view of the imported experiment's navigator in GeneSpring 13.0.

To be noted

- The imported experiment will always contain an All Aligned Reads list. If the All Aligned Reads list had been exported from the Strand NGS experiment, then the All Aligned Reads list in GeneSpring would be identical to the one in Strand NGS. If, however, this list had not been marked for export, the imported experiment in GeneSpring would still contain an All Aligned Reads list (Figures 4A and 4B). This would be the union of all the reads that have been imported, in which case the number of reads for this list in GeneSpring would be different from that in the Strand NGS experiment. When no read lists are selected for export, the All Aligned Reads list in GeneSpring would contain zero reads.
- All transferred read lists would be children of the All Aligned Read list (union of all reads).
- If the read lists chosen for transfer to GeneSpring have a parent-child relationship, that hierarchy is maintained in GeneSpring.
- If the read lists getting transferred are chosen so that the parent-child relationship is not maintained, there will not be a hierarchy in GeneSpring.
- In a DNA-Seq experiment, if a re-aligned read list is chosen for export, its immediate parent is implicitly exported too.
- Since all entity lists are exported by default, the hierarchy is retained upon import.

Table 1. List of objects available in various sequencing experiments and their export status.

Icon	Object	DNA-Seq	RNA-Seq	ChIP-Seq	Methyl-Seq	smallRNA-Seq
	Samples	☑	☑	☑	☑	☑
	Interpretations	☑	☑	■	☑	☑
	Read lists	☑	☑	☑	☑	☑
	Entity lists	☑	☑	☑	☑	☑
	Region lists	☑	☑	☑	☑	☑
	Cluster tree	☐	☐	☐	☐	☐
	Pathway list	☐	☐	☐	☐	☐
	SNP report node	☐	☐	■	■	■
	SNP reports	☑	☑	■	■	■
	SV node	☐	■	■	■	■
	SV region list	☑	■	■	■	■
	Copy number node	☐	■	■	■	■
	CNV pair node	☐	■	■	■	■
	Continuous data	☐	■	■	■	■
	CNV regions	☑	■	■	■	■
	Novel detection report	■	☑	■	■	☑
	Quantification node	■	☐	■	■	☐
	Gene fusion node	■	☐	■	■	■
	Gene fusion report	■	☑	■	■	■
	Motif node	■	■	☑	■	■
	Methylation node	■	■	■	☑	■
	DMC node	■	■	■	☑	■
	Active regions	■	■	■	■	☑

☑ Exported by default ☐ Cannot be exported
☑ Can be chosen for export ■ N/A

- Entity lists resulting from other operations, for example, *Translate Regions to Genes* and *SNP Effect Analysis*, are present as child lists of the original region list in Strand NGS. If the parent region list is not chosen for export; upon import, these entity lists become children of the All Aligned Reads list.

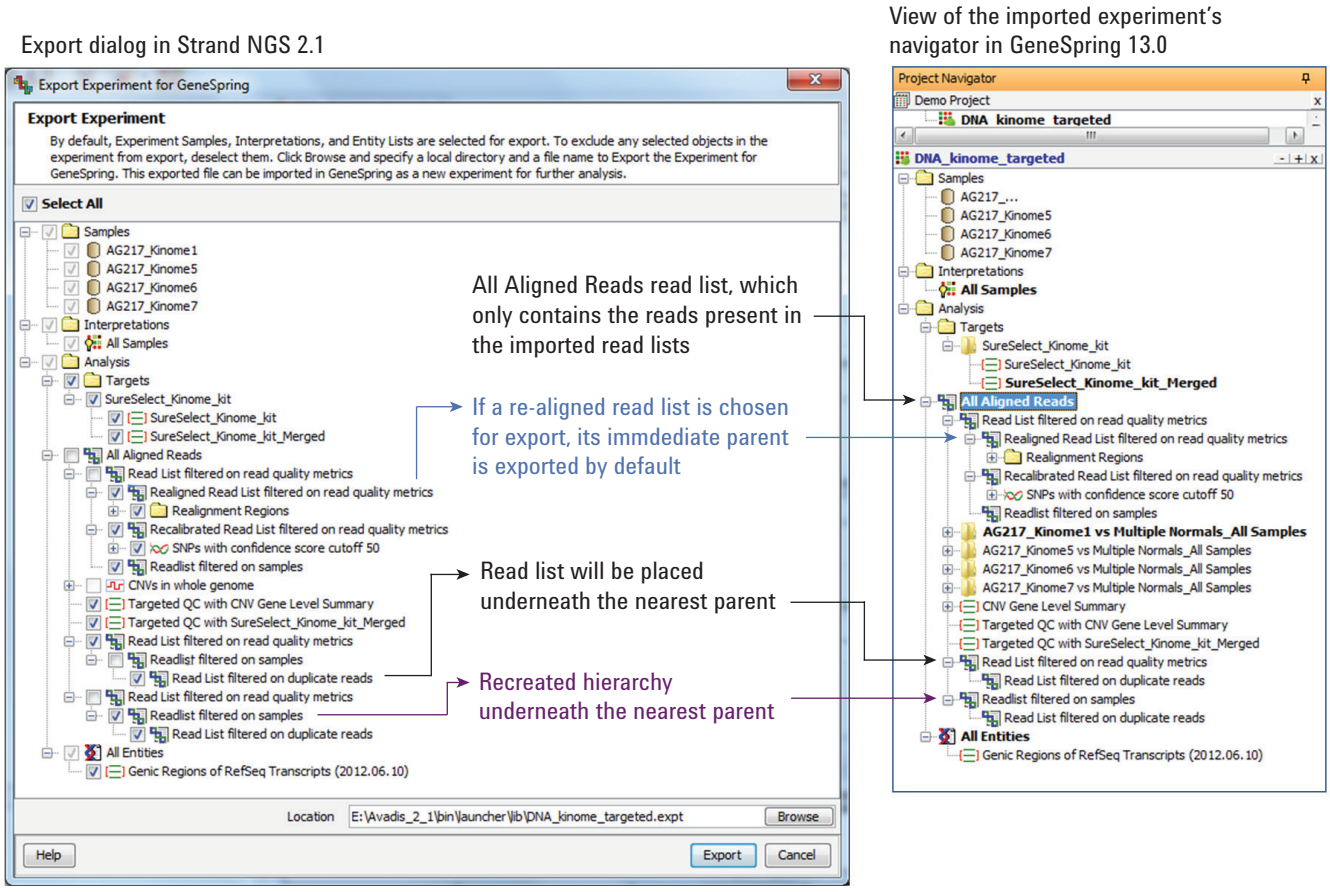

Figure 3. Maintenance of object hierarchy.

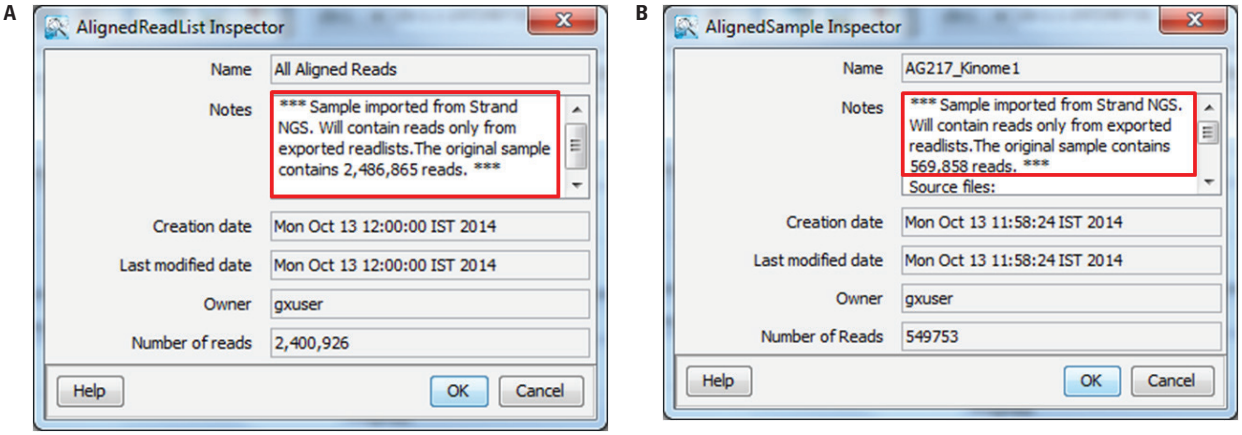

Figure 4. Inspecting All Aligned Reads and Samples. A) Read List Inspector - if All Aligned Reads is not exported, GeneSpring recreates an All Aligned Reads read list that is the union of all the reads present in the imported read lists. B) Sample inspector - original number of reads in each sample in the Strand NGS experiment is recorded in the Notes for each sample in the imported experiment in GeneSpring.

Read and Region List Attributes

All read and region list attributes are maintained when they are exported from Strand NGS, and imported into GeneSpring.

To be noted

- If, as the result of a filtering step, the mate M1 of a pair is filtered out and mate M2 is retained, upon import of this read list into GeneSpring, M2 will still retain the original mate status. M1, however, will be indicated in the Genome Browser as a read that is missing due to filtering, as seen in Figure 5.

Workflow

After an NGS experiment is imported into GeneSpring, the following operations, shown in Table 2, can be performed. For Pathway and Correlation analysis, creation of a Multi-Omics Analysis (MOA) experiment is a prerequisite. Visualization of reads in Genome Browser along with microarray data can be performed both at MOA and individual experiment levels. Additional visualizations such as scatter plots and Venn diagrams on entity lists can be performed only at the individual imported NGS experiment level.

Table 2. Support for various operations across different sequencing experiment types.

Analysis	DNA-Seq	RNA-Seq	ChIP-Seq	Methyl-Seq	smallRNA-Seq*
MOA creation	✓	✓	✗	✓	✓✓
A. Pathway	☐		✗	☐	
B. Entity correlation	✗		✗	✗	 ✗
Visualization in Genome Browser	✓	✓	✓	✓	✓✓
Cross-platform investigation	✓	✓	✓	✓	✓✓

* smallRNA-Seq is handled differently as it has two types of entities: active miRNAs and target genes (mRNAs). Both lists are supported for analysis in MOA. In each row, the first icon suggests the handling of miRNAs, and the second icon suggests the handling of target genes.

✓ Supported ☐ Pathway matches only  Entity correlation supported

✗ Not supported  Pathway matches and heatstrips

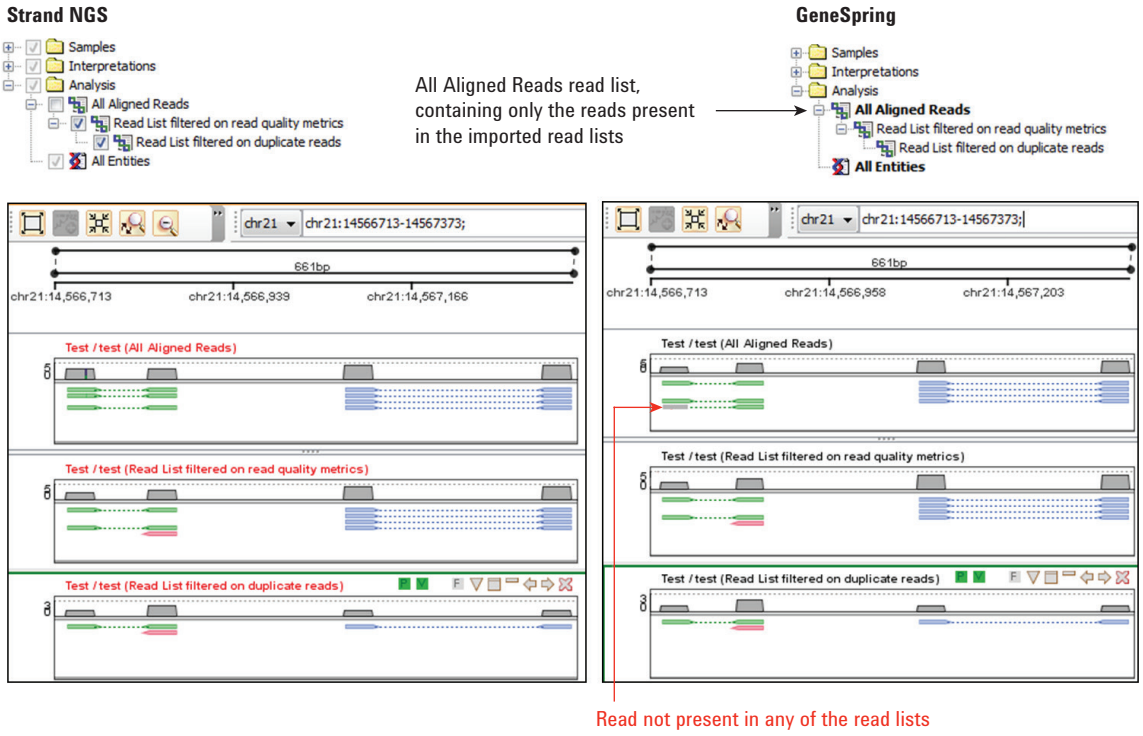

Figure 5. Missing mates (due to filtration step in Strand NGS) of read pairs are indicated as reads not present in the imported read lists. This will be seen only in the All Aligned Read list track of the Genome Browser and if the original All Aligned Read list was not exported.

Imported NGS Workflow Operations in GeneSpring

Experiment setup

After the import of the sequencing experiments into GeneSpring, new experimental grouping parameters can be added and interpretations can be created (except in the case of ChIP-Seq experiments, where the concept of interpretation is not relevant).

Multi-Omics Analysis

A new multi-omics experiment can be created using an imported sequencing experiment in combination with an array, mass spectrometry experiment, or two imported sequencing experiments. The option to create a MOA experiment is available in the workflow browser and the tool-bar. MOA enables pathway and correlation analyses to be performed on the sequencing data.

Pathway analysis

Entity lists from the sequencing analysis, and the other participating experiment (in MOA) can be used to identify enriched relevant pathways. Because of the ability to simultaneously visualize multiple biological phenomena, it becomes a powerful tool for elucidating cellular processes. For example, as shown in Figure 6, when yeast is grown under batch and chemostat mode, impact of structural variation[2] identified though DNA-sequencing, on the expression levels of the gene ADE13 (Adenylosuccinate lyase, involved in purine metabolism), assessed through microarrays and RNA-Sequencing, can be studied.

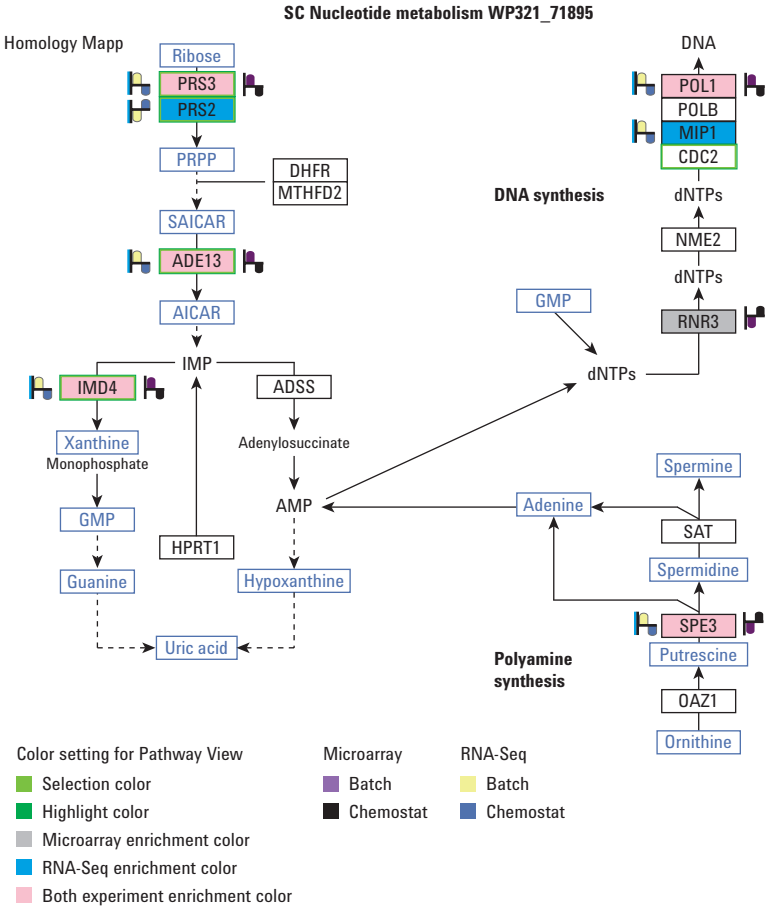

Figure 6. A pathway view depicting the expression levels of genes that show differential expression in yeast when grown under different conditions, namely batch and chemostat. The expression levels were independently measured using expression arrays and RNA sequencing, and are indicated by heat strips. The genes highlighted in green were found to have structural variations when studied using DNA sequencing. The pathway overlay shows the effect of structural variations on expression levels of genes that are key players for growth in batch and chemostat mode.

Correlation analysis

The correlation framework in GeneSpring supports correlation between molecules studied by a single omics platform, or between two different omics platforms. Thus, entities from a sequencing experiment can be used for multi-omics entity level correlation. For example, in the study published by Hesse *et al.*[3] shown in Figure 7, entity level correlation analysis performed on the miRNA and mRNA identified to be differentially expressed by SmallRNA-Seq and microarray analysis, respectively, showed a negative correlation of let-7 family members with PIGU.

Visualization

Informative patterns of genetic and epigenetic regulation often become apparent with appropriate and powerful visualizations. These include viewing together sequencing and array data in Genome Browser, Scatter plot, and so forth. This enables the selection of interesting entities from one experiment in various views such as a Venn diagram, scatter plot, and a heat map of other experiments. Figures 8 and 9 show visualization examples of methylation and gene expression levels that enable identification of relationships between the two regulatory processes. The promoter and gene body methylation status in a triple negative cell line[4], as assessed using bi-sulfite sequencing was analyzed using Strand NGS. Differentially expressed genes from tissue samples[5] were examined in GeneSpring in context of genes identified to be exhibiting hypermethylation in Strand NGS.

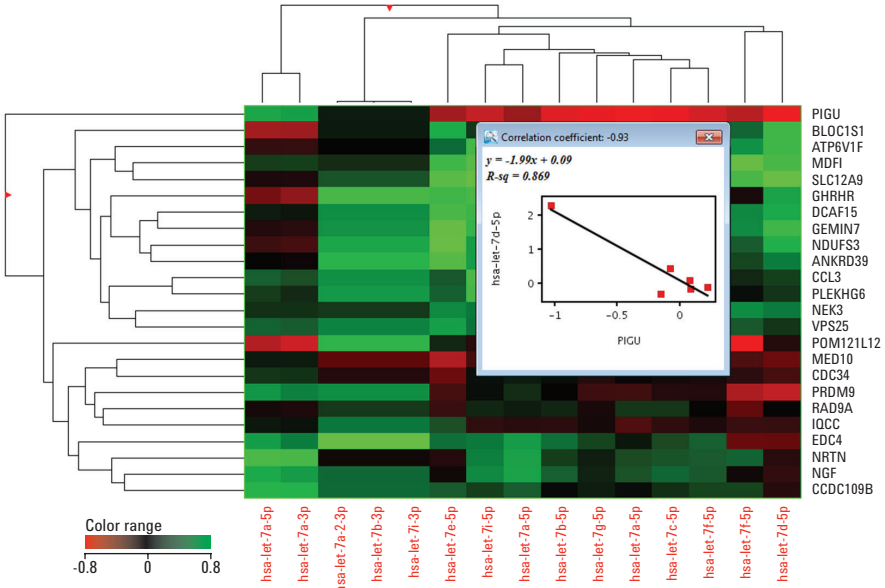

Figure 7. A heatmap depicting the correlation between the let-7 family of miRNAs and their known gene targets in normal mammary cells, in which ATM has been deleted. From this heatmap, it can easily be inferred that most of the members of the let-7 family are negatively correlated with PIGU, a protein with similarity to yeast cell division cycle protein CDC91.

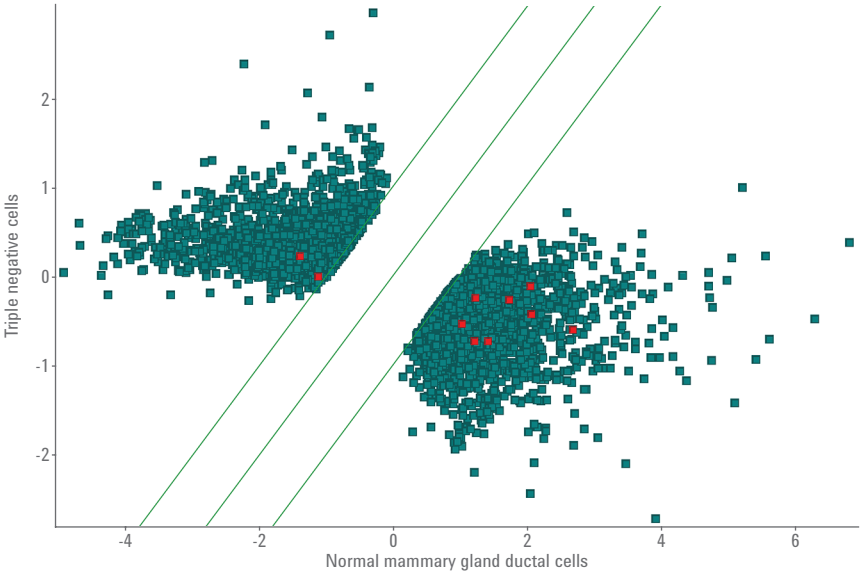

Figure 8. A scatter plot depicting the genes found to be showing expression changes > 2 fold between triple negative cells and normal mammary cells when studied using microarray. The genes highlighted in red were found to be hyper-methylated in MDA-MB-231, a triple negative cell line. The scatter plot was launched in the microarray experiment on the entity list showing fold change > 2.0. Genes with promoter methylation were highlighted in this scatter plot by making it the active entity list.

The genes with hypermethylated promoters were identified by performing following sequence of operations in Strand NGS:

1. Methylation detection.
2. Identification of differentially methylated cytosines (DMCs) between MDA_MB-231 and the normal mammary cells.
3. DMCs were used to determine the Differentially Methylated Regions or the DMRs.
4. A subset of DMRs present in the promoter regions were obtained by comparing DMRs with the genome wide genic and promoter regions (using compare region list operations).
5. Promoter DMRs were translated to genes (translate regions to genes functionality in Strand NGS).

Subsequent to Step 5, the Methyl-Seq experiment with methylated cytosines, DMC and DMR region lists and all entity lists were exported from Strand NGS as an .expt file and imported into GeneSpring 13.0. Genes with methylated promoters were then used for overlap through a Venn diagram, with those genes that were differentially expressed (differential expression analysis of microarrays in GeneSpring) to identify genes affected by promoter methylation.

As seen in the scatter plot in Figure 8, most of the genes with methylated promoters in the triple negative cell line show reduced expression in patient tumors. However, a handful of genes do show higher expression levels in spite of promoter methylation.

Figure 9. An elastic Genome Browser view depicting a gene showing promoter hyper-methylation in cell lines and decreased expression in tissue samples. The first two tracks show locus-specific methylation levels in promoters identified to be DMRs across normal and tumor cells, respectively. The third track pinpoints the DMCs while the fourth and the fifth tracks display open reading frame and expression fold change, respectively. The expression levels were measured using microarrays on tissue samples. The Genome Browser view is from GeneSpring, with the first three tracks showing data from the imported methyl-seq experiment, while the fifth track shows data from the microarray experiment.

Interactive exploration (cross-platform investigation)

Transcription factors (TFs) are important proteins involved in regulating gene transcription. Chromatin-Immunoprecipitation-sequencing (ChIP-Seq) is a method used to identify genomic regions in target genes bound by TFs corresponding to TF binding sites (TFBSs). Examining these targets in the context of their differential expression allows one to explore the pathways involved in a given cellular process. Galhardo et al.[6] studied the dual regulation by transcription factors and microRNA of the genes involved in adipocyte differentiation. LXR is a transcription factor involved in fat metabolism, and genes regulated by LXR are targeted by drugs for cardiovascular conditions. In this study[6], genes whose promoters were bound by the transcription factor LXR were identified using ChIP-Seq. The same group also conducted microarray experiments to understand the time course of adipocyte differentiation, and to identify genes involved in fat metabolism in the presence and absence of miR-29a. Integrative analysis of the two microarray and one ChIP-Seq experiment helped establish interaction networks for five genes that were differentially expressed during adipocyte differentiation, were regulated by miR-29a, and had LXR binding to their promoters, as shown in Figures 10A and B.

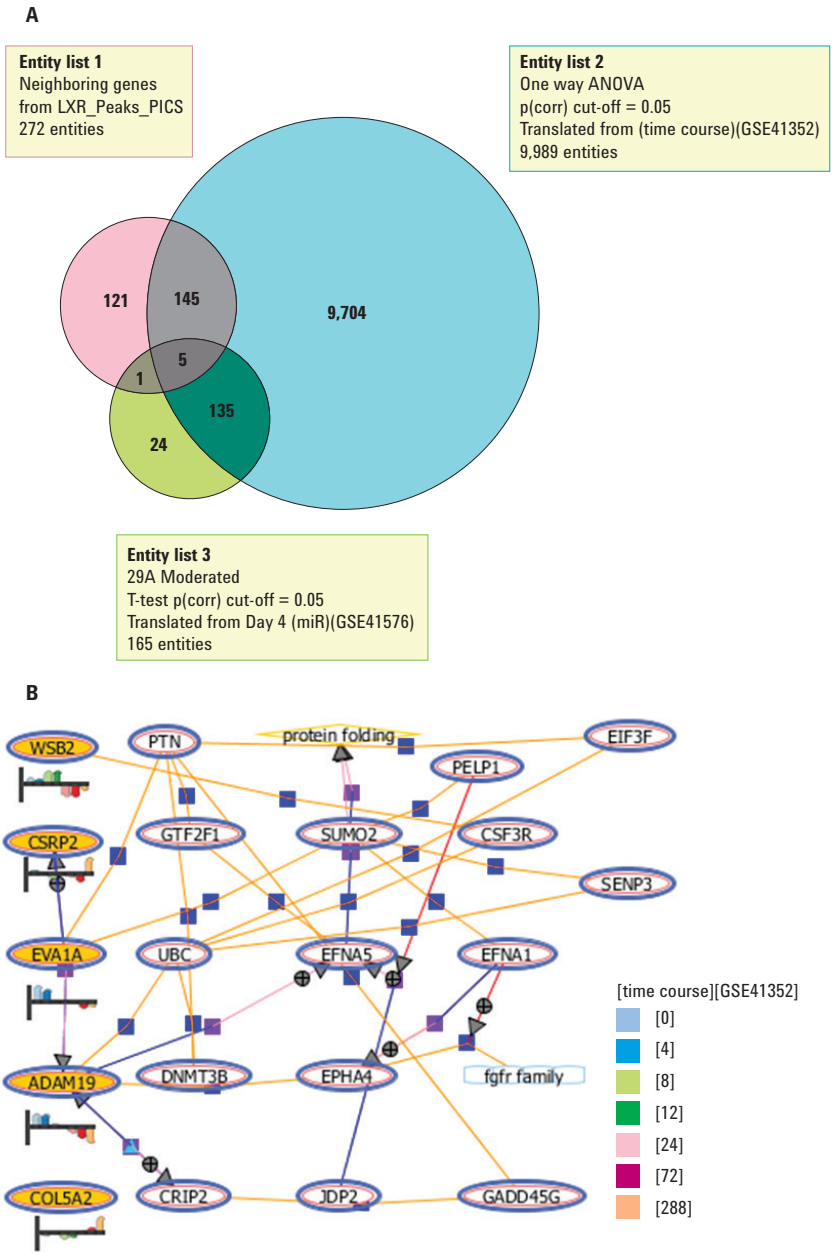

Figure 10. A) A Venn diagram showing the overlap between genes known to be differentially expressed during the course of adipocyte differentiation, genes regulated by miR-29a, and genes whose promoters are bound by transcription factor LXR. B) Interactions between the five overlapping genes discovered using Natural Language Processing (NLP).

Conclusion

Complexities of cellular dynamics can be best understood when complementary information from various sources are interpreted in relation to each other rather than in isolation. Thus, while a microarray study, on the basis of differential expression, can provide information regarding the genes that might be involved in a process, a DNA-sequencing of the same samples can determine if the same genes have copy number aberrations. Similarly, ChIP-Seq would provide data connecting differential expression to differential regulation by a transcription factor. All this information, when processed appropriately and viewed together, allows the user to discern levels of regulation whereby some members of a pathway could be up-regulated due to gene amplification while others show higher expression due to promoter regulation. To leverage maximum potential from such an approach, GeneSpring 13.0 has facilitated the import of sequencing experiments so that they can be analyzed, viewed, and interpreted in a multi-omics context.

References

1. Rhodes, D. R; Chinnaiyan, A. M. Integrative analysis of the cancer transcriptome. *Nature Genetics* **2005**, *37*, S31-S37.

2. Nookaew, I; *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res*. **2012**, *40(20)*, pp 10084-10097.

3. Hesse, J. E; *et al.* Genome-Wide Small RNA Sequencing and Gene Expression Analysis Reveals a microRNA Profile of Cancer Susceptibility in ATM-Deficient Human Mammary Epithelial Cells. *PLoS ONE* **2013**, *8(5)*: e64779.

4. Hodges, E; *et al.* High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res*. **2009**, *19(9)*, pp 1593-605.

5. Komatsu, M; *et al.* Molecular features of triple negative breast cancer cells by genome-wide gene expression profiling analysis. *Int. J. Oncol.* **2013**, *42(2)*, pp 478-506.

6. Galhardo, M; *et al.* Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. *Nucleic Acids Res*. **2014**, *42(3)*, pp 1474-1496.

www.agilent.com/chem

For Research Use Only.
Not for use in diagnostic procedures.

PR7000-0096

Published in the USA, January 4, 2016
5991-5528EN

Agilent Technologies