Improving Outcomes Through Enhanced Data Analytics and Artificial Intelligence

Part 1: Many Roads to Reproducibly Wrong Results—Distinguishing the Truth Using ReSurfX::vysen and Other Analytic Workflows

Automated, complex, and sophisticated knowledge extraction in the form of artificial intelligence (AI) is maturing rapidly for application to every sector including the traditionally recalcitrant healthcare and life sciences (Pharma and Biotech). With ever-increasing capabilities in big data generation and utilization, adopting this digital revolution is critical for organizations to be successful.

This four-part feature discusses: (i) testing analytics workflows with the right data and metrics, (ii) what happens to your favorite analytics' accuracy and ability to derive novel insights when it comes to real-world big data and unknown error properties, (iii) the reality of the knowledge repositories upon which our AI engines rely, and (iv) if machine learning is the cure for limitation of statistics and enterprise productivity.

In this first installment, we evaluate a variety of widely used analytic approaches in comparison to ReSurfX::vysen that utilizes the novel Adaptive Hypersurface Technology (AHT) on gene expression and differential expression analysis from sequencing and microarrays – often the first step of their utilization for variety of drug discovery, development and disease diagnosis, and high-cost drug sales and patient care delivery.

Current Problems

Current analytic approaches specifically designed to data from a technology platform have too many errors. The resultant errors (i) waste resources spent on analysis and downstream uses, and (ii) provide confounding false information that decreases innovation and ability to derive novel insights.

Root causes of analytical failure and significant reduction in ROI in data-driven business goals

The advantage of the ability to collect, store, and process large volumes of data comes with an inherent disadvantage as well. Many classical analytics and machine learning tools fail in the presence of a strong dependency on fundamental statistical metrics. It should also be noted that with increasing data volume, what was previously considered a small proportion of error is now a very large amount of error in a workflow. For example, an error rate of 1/1000 in a petabyte of data results in a terabyte of error in the workflow. Statistics-based data or error modeling makes assumptions that fail more often in big data due to non-uniformity, so we highlight and refer to that property of error in big data explicitly as 'errors are unpredictable'. AHT is built to be robust against unpredictable error by avoiding explicit error modeling.

> An error rate of 1/1000 in a petabyte of data results in a terabyte of error in the workflow.

A salient highlight is the strong sentiment against misuse of p-value based statistics (most other statistics are also built on that concept) from American Statistical Institute [ASA-guidance] to popular media [e.g., John Oliver, Vox].

Statistics-based analytics fail primarily due to (i) too many hypothesis tests in large volumes of data (big data) – termed 'multiple testing problem', and (ii) often the sample sizes used are small, this problem is termed 'lack of statistical power' to achieve high confidence results. The small sample size problem is more obvious in discovery or early stage exploration of even late stage development or high value medical care deployment (e.g., rare diseases, a disease or drug applicable to a subset of population classified to have a disease in broader terms – cancer is an example of latter where significant progress is being made).

Even in studies that involve significantly higher number of subjects, well-known statistical techniques like Bayes, or modern techniques, such as Deep Learning, require significantly large amounts of data to be reliable. Thus, a different class of sample size exist when using AI/ML – commonly known as the 'learning set'.

Advantages of Adaptive Hypersurface Technology (AHT) and ReSurfX::vysen product

Adaptive Hypersurface Technology, invented by ReSurfX team[2], differs from most analytic approaches in the sense that it is a data-source agnostic machine learning approach that primarily optimizes a 'number' of input parameters that help evaluate an outcome of interest from given data. Thus, (i) the input parameters are not pre-fixed, as in many feature selection approaches, and (ii) the 'number' of input as opposed to a specific set of input-parameters, gives an opportunity to (a) determine differences closer in data space, (b) always lean on multiple inputs, even in cases where data-science specialists find ways to summarize combinations of input, increases robustness, (c) give combinatorial possibilities to determine the output and outcome of interest from the input data, thus adding a component of inherent personalization to analytics workflows, and (d) the data-source agnostic nature of AHT lends itself to many applications with lesser need for customization. It is naturally integrative with sophisticated and enterprise-specific workflows currently in use, blending with the power of other technologies and specialized knowledge of subject matter experts.

Adaptive Hypersurface Technology is a data-source agnostic machine learning approach that primarily optimizes a 'number' of input parameters that help evaluate an outcome of interest from given data.

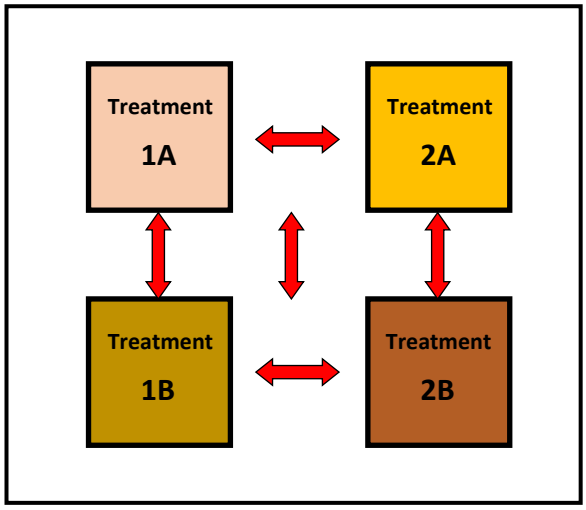
Products and solutions leveraging AHT range from end-to-end applications in Pharma (drug-development) and healthcare enterprises to provide deep expert support for disease discovery and management, clinical trials, and delivery of high-cost patient care products. Most of the results presented below were generated using our enterprise SaaS product, ReSurfX::vysen.

The in-market enterprise grade SaaS product, ReSurfX::vysen, showcases the power of AHT as a customized solution for RNASeq based and Affymetrix GeneChip based microarray measurements – as these two form most of the gene expression measurement data in the world. Other uses and specialized advantages are highlighted at https://resurfx.com/about and https://resurfx.com/vysdom.

Box 1

1. All other workflows for data from GeneChip and RNASeq use summary measures (i.e. reduce multiple independent measurements into single value for each gene), whereas AHT used in ReSurfX::vysen does analysis using the multiple measures directly.
2. All other workflows necessitate the use of subjective user specified thresholds like p-value, false discovery rate, etc., whereas vysen gives differences as 0 and 1 with a confidence measure.
3. All other workflows compared here use log2 transformation while ReSurfX::vysen does analysis on non-transformed data.
4. For GeneChip, current workflows use perfect match (PM), whereas vysen uses perfect match - mismatch (PM-MM) as the signal.

Fig. 1—Reference Study Design



Gene expression data: Treatment A, a mixture of 10 pooled cancer cell lines, and Treatment B, a collection of different regions of the human brain, were collected from two sites Site1 and Site2, called 1A/1B and 2A/2B respectively, and used for a variety of intra- and inter-treatment analysis.

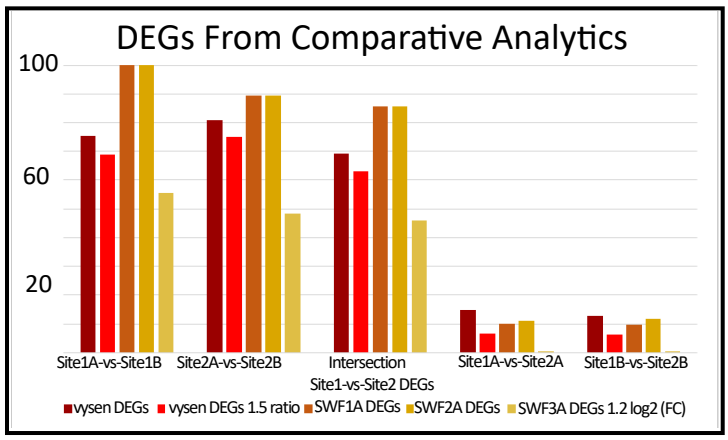
Objective of this whitepaper and dataset used

This whitepaper leverages widely used data generated by sequencing and microarray quality control consortium[3] using Illumina instruments (small fragment RNASeq) and GeneChip microarray platforms data from two independent sites (of 5 or more sites from which the data are available) generated for two treatment comparisons (A-vs-B) with each treatment having 5 technical replicates. The simplest study design is depicted in Fig. 1 and the results are interchangeably referred to as 'reference results' or '5-mer analysis' results are indicated in Fig. 2. We compared performance of vysen with two widely used workflows for GeneChip and RNASeq, and an additional workflow from a commercial product for GeneChip[4]. We refer to each of these analytic workflows as a numbered SWF (SWF1A, SWF2A etc.) - Box 2. Some key differences between ReSurfX::vysen and the SWFs are highlighted in Box 1.

Box 2

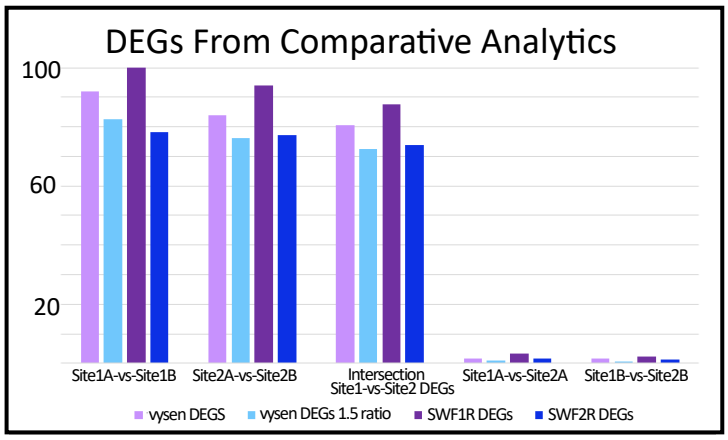
Input: GeneChip CEL files	**Input:** Aligned BAM files (using HISAT[9])
READ DATA & NORMALIZE **SWF1A, SWF2A, SWF3A:** Affy package[6] & rma[6] **vysen:** 'Extraction' – i.e., data format conversion, proprietary AHT normalization	**READ DATA & NORMALIZE** **SWF1R, SWF2R:** featureCounts[10] & negative binomial[11] **vysen:** 'Extraction' – i.e., data format conversion, proprietary AHT normalization[12]
COMPARATIVE ANALYTICS **SWF1A:** limma[7] (eBayes) **SWF2A:** t-test[8] **SWF3A:** t-test[8] **vysen:** AHT (multi-parametric)	**COMPARATIVE ANALYTICS** **SWF1R:** glm[13], (related to limma, uses eBayes) **SWF2R:** edgeR[11](Fishers Exact Test) **vysen:** AHT (multi-parametric)
DEG SELECTION **SWF1A,** Amean > 25th, percentile, p < 0.01, **SWF2A:** fdr (Benjamini-Hochberg) < 0.1, ratio >= 1.5 **SWF3A:** signal values > 25th percentile, p < 0.05, fold change > 1.2 **vysen:** no external selection (AHT provides 0 or 1)	**DEG SELECTION** **SWF1R,** Amean > 25th, percentile, p < 0.01, **SWF2R:** fdr (Benjamini-Hochberg) < 0.1, ratio >= 1.5, 2/5 samples > 10 counts & cpm > 0.2 (per gene) **vysen:** no external selection (AHT provides 0 or 1)

Fig. 2 — The reference analysis results comparing gene expression in treatments A and B of the microarray data



A-vs-B comparisons using all the 5 technical replicates at each site to identify differentially expressed genes (DEGs) using vysen, two widely used workflows (SWF1A and SWF2B) and a commercial workflow (SWF3A). The most number of DEGs identified (19,375) was set as 100 and the rest are indicated as proportional percentage values. **It should be noted that the intra-treatments (A-vs-A, and B-vs-B between the sites were less than 10%).**

Fig. 3— The reference analysis results comparing gene expression in treatments A and B of the RNA-seq data



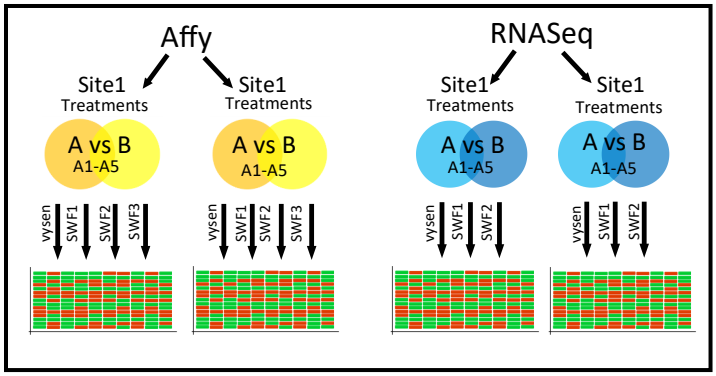
A-vs-B comparisons using all the 5 technical replicates at each site to identify differentially expressed genes (DEGs) using vysen, two widely used workflows (SWF1R and SWF2R). The most number of DEGs identified (25,112) was set as 100 and the rest are indicated as proportional percentage values. **It should be noted that the intra-treatments (A-vsA, and B-vs-B between the sites were less than 10%).**

Fig. 4— Generating 45 comparisons from each 5-replicate treatment

	1,2,3	1,2,4	1,2,5	1,3,4	1,3,5	1,4,5	2,3,4	2,3,5	2,4,5	3,4,5
1,2,3	X									
1,2,4	2	X								
1,2,5	2	2	X							
1,3,4	2	2	1	X						
1,3,5	2	1	2	2	X					
1,4,5	1	2	2	2	2	X				
2,3,4	2	2	1	2	1	1	X			
2,3,5	2	1	2	1	1	1	2	X		
2,4,5	1	2	2	1	1	2	2	2	X	
3,4,5	1	1	1	2	2	2	2	1	2	X

All possible 3-mer combinations from each treatment (A or B from Site1 or Site2) were generated and compared as indicated above. The numbers inside the cells indicate the number of common samples in that comparison.

Fig 5—Reference and subset comparisons yield various metrics of performance for vysen and other analytic workflows



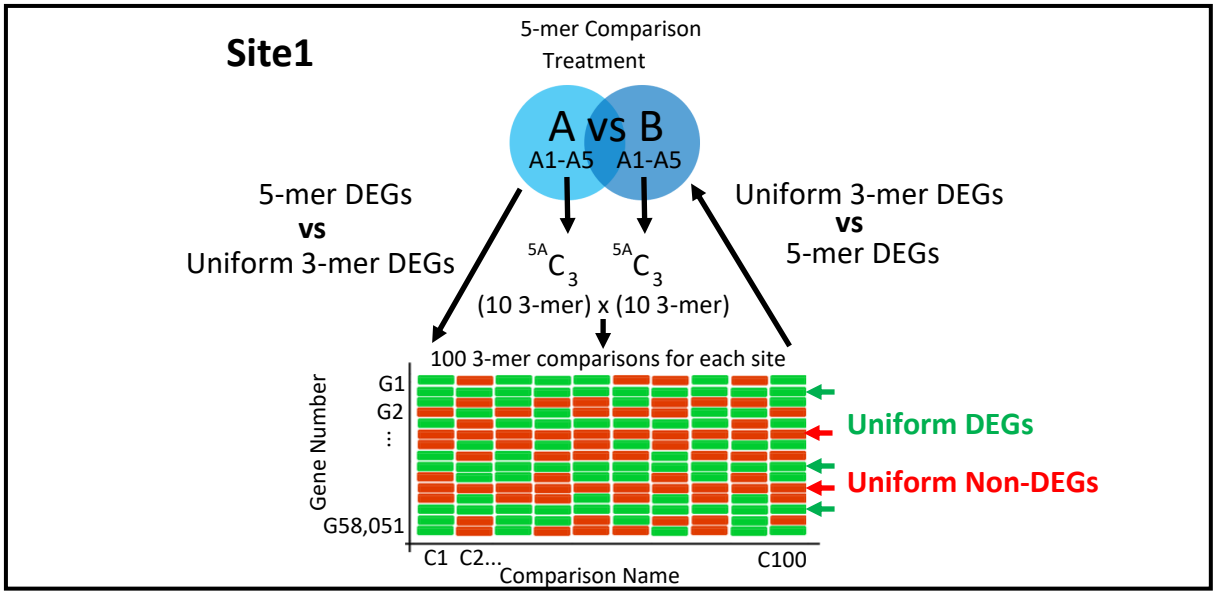
The subset inter-treatment study provided a large number of comparisons (100 per Site per workflow). Those subset comparisons and the reference (5-replicate) comparisons shown in Fig.4 were applied to vysen and each of the other workflows to derive comparison metrics of reproducibility, accuracy, precision and recall.

Additionally, we utilized this data to generate a large number of comparisons and many different metrics teasing out various metrics of analytics performance (accuracy, reproducibility, precision, recall etc.). One of the key conclusions is that the most consistent performance was obtained with vysen and the commercial SWF (SWF3), with the latter consistently detecting about 50% of differentially expressed genes (DEGs) in data presented here and large number of other studies.

A Subset study design to evaluate metrics of performance of different analytics approaches

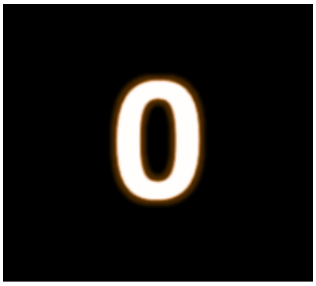
A key aspect of our study design is to derive all possible '3-mer' (3 replicates of 5 replicate) from the 5 replicates for each treatment (A or B) for each technology and site ('subset comparison'). We thus were able to generate results from a lot of comparisons, and in comparison with the master results, the reference (5 replicate) original study design derive to evaluate a variety of metrics of analytic performance. This study design is represented in Fig. 4, Fig. 5 and Fig. 6.

Fig 6—A master subset comparison using reference and 100 3-replicate intra-treatment comparisons



Shown is a depiction of generation of 100 3-replicate comparisons. Metrics of performance such as reproducibility, recall, precision etc., were derived by using these 100 comparisons as well as comparing them to the reference (5-replicate) comparison for each site and each workflow.

Fig. 7—Zero false positives were identified in intra-treatment analysis



vysen and in all the workflows with tuned analysis parameters used here identified ZERO DEGs were observed in the 180 3-mer intra-treatment (A-vs-A or B-vs-B) analysis spanning over 5 million DEG evaluations per technology per workflow. Exceptions were: 45 and 92 DEGs (false positives) were observed over 5 million calls of SWF1R and SWF2R.

ZERO false positives in subset study design for intra-treatment analysis

One of the astonishing observations of this study is that ZERO false positives can be obtained using vysen as well as a variety of analytical workflows with little tuning of analysis parameters (Fig. 7).

The first metric we derived using this study design was to look at the number of DEGs (theoretically idealized FALSE POSITIVES) when comparing A-vs-A and B-vs-B for each site for each technology. This comparison design is indicated in Fig. 3 and yielded ZERO false positives across all the 45 comparisons for each sample for each technology. In other words, we got ZERO false positives across 180 comparisons representing over 10 million differential expression calls. Because finding ZERO false positives was unprecedented and previously unreported (probably because this study design was not explored with this data) we used it to tune metrics to use with the different SWFs. The result of the tuned parameters in SWFs, naturally adhering to the purpose of the tune, resulted in ZERO or very low (45 and 92 false positive DEGs were identified by SWF1R and SWF2R workflows of RNASeq, respectively) false positive DEGs in over 5 million differential expression calls for each data collection platform (Fig. 7).

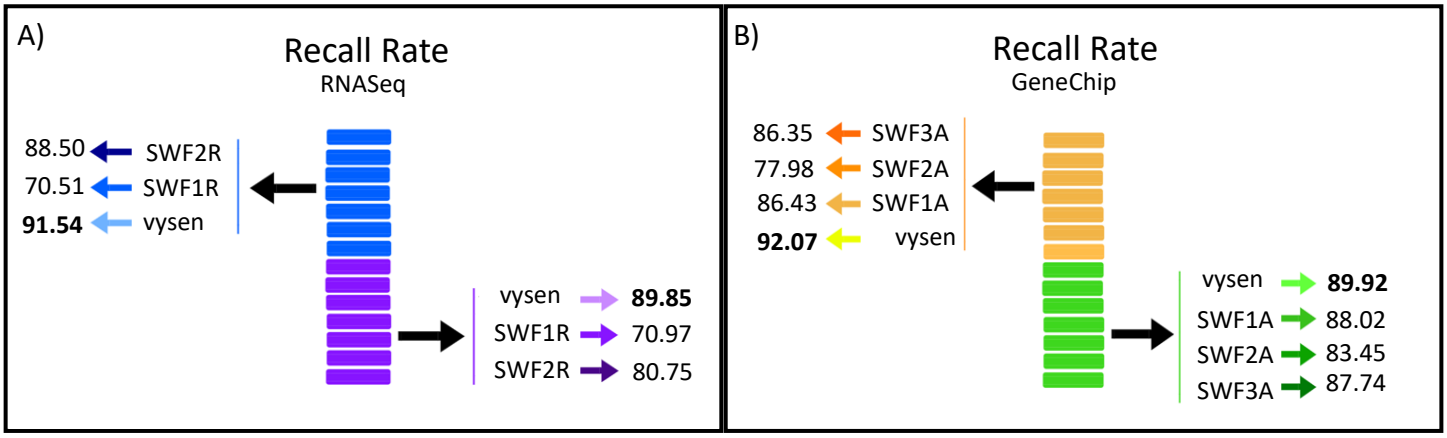
The above comparison helped evaluate comparative analytics step of the workflow without confounding batch-effect correction errors. However, the traditional eFDR (empirical False Discovery Rate) across sites done by normalizing each of the relevant 10 samples together (batch-effect correction) indicates that the error rate was less than 10% with vysen having the least errors. In the previous statement, application of a ratio threshold of 1.5 to ReSurfX::vysen as well to get equivalent value when comparing to SWFs.

Use of tuned input thresholds in SWFs to compare analytics performance against ReSurfX::vysen

We used the metrics represented there to evaluate:

i. Recall rate of 'Uniform DEGs' (subset of DEGs those were uniform across the 100 3-mer comparisons between the two samples A-vs-B – i.e., genes called DEGs with 100% precision) when the number of replicates was reduced to three, as percentage of reference (5-replicate) analysis (Fig. 8A & Fig. 8B). ReSurfX::vysen had the highest recall rate of all the workflows tested here.

Fig 8A & B—Recall rate of DEGs with 100% precision in 3-mer comparison compared to reference



Shown are **recall rates** as percentage of DEGs in the **reference** (5-replicate) comparison that were also present as DEGs with **100% precision** in the 100 subset (3-mer) inter-treatment comparisons. **(A)** shows results for RNASeq sequencing and **(B)** shows result for GeneChip microarray. This represents the 5-mer DEGs vs ' in Fig. 5.

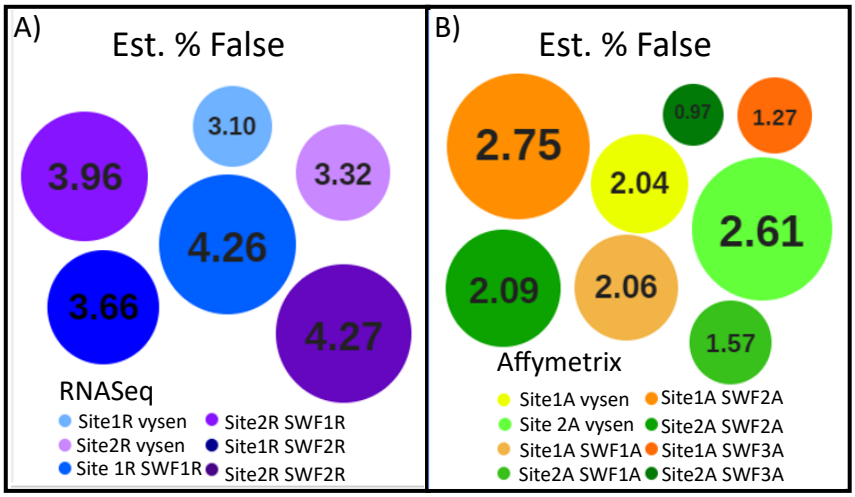
ii. Precision across all the 100 subset (3-mer) comparison differential expression calls was measured indirectly as estimated errors – i.e., using all DEGs and Non-DEGs to calculate the percentage error Fig. 9A & Fig. 9B. ReSurfX::vysen had the best precision in RNASeq analysis and SWF3A had the best precision for GeneChip analysis. However, SWF3A had the least (significantly lower) DEGs of all workflows.

iii. Accuracy (and spread) in the 100 subsets (3-mer) comparison when 'DEGs from each of comparison' are individually compared to the DEGs in the 'reference comparison' (as shown in Fig. 10) using the same analytics workflow (Table 1) – vysen had most variation (increase) in total number of DEGs when the number of replicates are reduced. However, with well-known results from statistics-based analytics, this result may not translate to most real-world data. Checking the converse, i.e. the DEGs of reference analysis to each of the subset analysis, vysen had the DEGs of former is contained in the latter better than any other workflow.

iv. Estimated error rate of DEGs and non-DEGs with 100% precision in subset (3-mer) inter-sample (A-vs-B) comparisons not present in the equivalent reference comparison – called 'Uniform 3-mer DEGs-vs-5-mer DEGs' in Fig. 6, which indicates:

- For GeneChip, the error rate (i.e. the differential calls with 100% precision not having same call in the 'reference (5-mer) analysis') of the reductive model approach is ZERO. This is indicative of the expected results, as the reductive modeling was originally developed with technical replicates of this nature. It is interesting that though vysen was not developed or trained using this class of data, it nevertheless gives an error rate of ZERO.
- In the case of RNASeq, the error rates shown in Fig. 11 exposes the flaw in SWF2R (which has the most error) due to the fact that Fisher's Exact Test applied in edgeR is not amenable to parametric analysis. This data, together with SWF1R having the least recall in Fig. 8, indicates that SWF1R likely has more false negatives and SWF2R has more false positives.

Fig. 9A & 9B—Measure of precision of DEG and Non-DEG calls in the 100 subset comparisons



Precision is represented as estimated false positives in the 100 3-mer inter-treatment comparisons in over 10 million differential expression calls. Estimated false positives were calculated by summing the number of times a gene was called Non-DEG (considered wrong DEG calls) when the total DEG calls for that gene were 50-100 DEG across the 100 subset comparisons and number of times genes were called DEGs calls (considered wrong Non-DEG calls) when the number of DEGs were 1- 50 across the 100 subset comparisons. This was done for each of the two technologies RNASeq and GeneChip microarray.

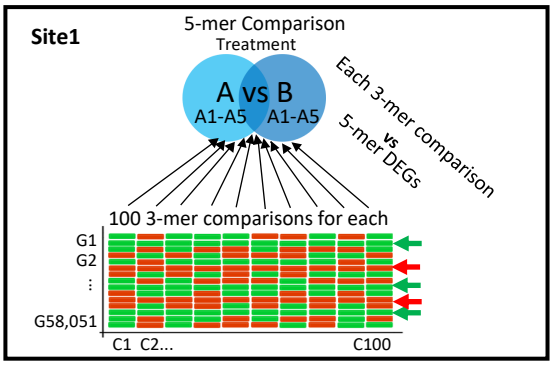
Table 1—Accuracy (and spread) of DEG calls when the number of replicates were reduced

Affy		3-mer DEGs range across 100 compares	3-mer DEGs median	overlap percentage range	overlap percentage median
vysen 5-vs-3mers	Site1A	15284-15871	15639	90.39-93.74	91.99
vysen 5-vs-3mers	Site2A	16402-17475	17178	87.81-93.02	89.75
SWF1A 5-vs-3mers	Site1A	19312-19493	19409	83.99-85.00	84.54
SWF1A 5-vs-3mers	Site2A	17281-17423	17347	93.34-94.22	93.91
SWF2A 5-vs-3mers	Site1A	18085-18873	18439	96.45-97.77	97.21
SWF2A 5-vs-3mers	Site2A	16810-17195	16968	96.41-97.60	97.27
SWF3A 5-vs-3mers	Site1A	10557-11020	10806	94.94-97.38	96.48
SWF3A 5-vs-3mers	Site2A	9328-9577	9440	95.74-97.61	97.02
RNASeq		**3-mer DEGs range across 100 compares**	**3-mer DEGs median**	**overlap percentage range**	**overlap percentage median**
vysen 5-vs-3mers	Site1R	23075-25223	24298	89.19-95.74	92.40
vysen 5-vs-3mers	Site2R	21103-23229	22258	88.02-98.12	91.74
SWF1R 5-vs-3mers	Site1R	20089-24788	21709	94.28-99.61	97.60
SWF1R 5-vs-3mers	Site2R	19555-23180	19987	94.55-99.25	98.27
SWF2R 5-vs-3mers	Site1R	22389-25425	23814	78.96-82.00	76.34
SWF2R 5-vs-3mers	Site2R	20419-23781	20919	79.47-83.17	82.05

The tables represent the number of DEGs in 'each of the subset comparisons' - range and median, as well as the percentage overlap with the reference (5-mer) comparison. (A) presents the data for GenChip and (B) presents the data for RNA-seq. It should be noted that vysen identified higher DEGs in the subset comparisons hence comparing the reference DEGs to individual subset DEGs yielded similar or better percentage overlap.

These results indicate that much of the Affy analytics are overfitted using data of this nature. As such, it is apparent that vysen has superior performance when we consider the number of probe-pairs AHT considers relevant to make a differential call decision as shown in Fig. 13A & 13B.

Fig. 10—Assessing the effect of reducing replicates in comparisons on accuracy



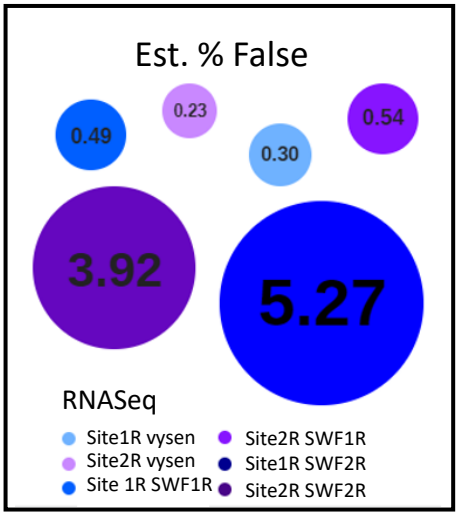
Each of the subset (3-replicate) comparisons (in Fig. 5) was compared to the Reference (5-replicate) comparison. This measures the estimated accuracy (and spread) over 100 comparisons - i.e., effect of reducing replicates. This analysis was done for each of the workflows as shown in Fig. 5.

Finally, it shows that the variety of metrics fail to expose most shortcomings in any of the analytics used, likely because the data and error in the technical replicates follows most of the statistical assumptions, reductive modeling of GeneChip data, eBayes, linear modeling etc., in the parametric comparative analytics except the obvious flaw in SWF2R, and likely higher proportion of false negatives in SWF1R. All the data indicate that vysen is the superior workflow of the three tested for RNASeq data. In addition, vysen has the best recall rate in GeneChip data analysis, which is an indicator of performance of both the reference comparison as well as the 3-replicate comparisons.

ReSurfX::vysen performs superior to all other workflows compared here

The data overlap from different workflows is also presented as Venn diagrams in Fig. 12A, B, C, & D. SWF2A had near complete overlap with SWF1A for the data analyzed here. Importantly, the overlaps and unique DEGs between the workflows acts as a powerful information to study the quality of the different workflows. Here we show information using the GeneChip data that provides important data on the quality of the workflows. The boxplot in Fig. 13A & B show that by the internal number of probes AHT uses ReSurfX::vysen data must be more accurate based on the number of fragments of each gene used to make the DEG or non-DEG call. Additional information supporting the fact that the vysen-specific differences cannot be captured using various machine learning tools will be presented in **whitepaper #4**.

Fig. 11—Effect of reducing replicates in comparisons on accuracy and reproducibility



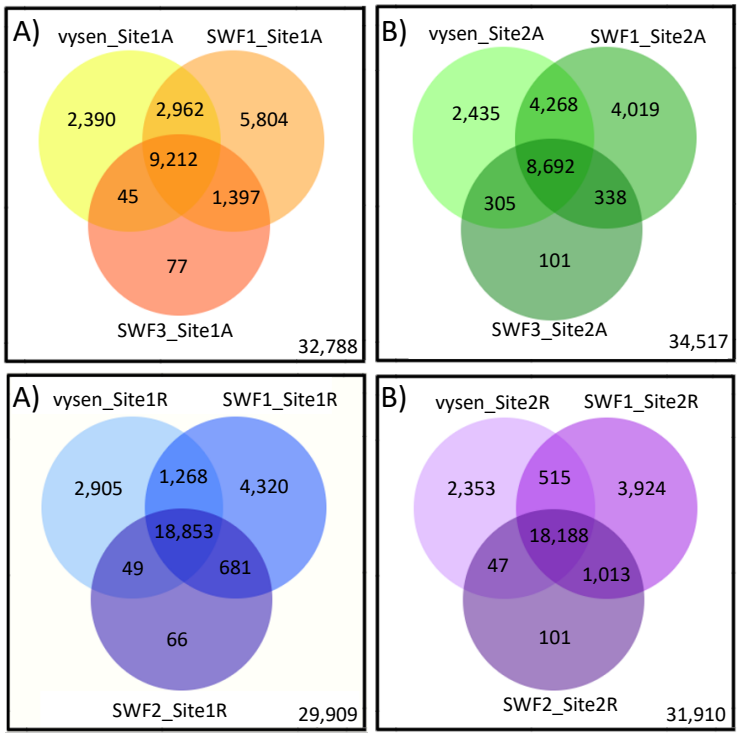
Proportion of DEGs and Non-DEGs with 100% precision in subset (3-mer) analysis that has equivalent call in the reference (5-mer) comparison. This is the 'Uniform 3-mer DEGs-vs-5-mer DEGs' shown in Fig. 5

The boxplot shown in Fig. 13 displays the number of individual probe-pairs (independent measures of a gene) used in making the decision to call the subsets of genes shown as DEG or Non-DEG. The data there for DEGs uniquely found by ReSurfX::vysen or eliminated as DEGs compared to SWF1A in GeneChip are the most informative as the unique DEGs of vysen are false positives if SWF1A and unique Non-DEGs of vysen are true positives of SWF1A. The result clearly indicates that the adaptation of AHT used in ReSurfX::vysen is providing superior results. As can be seen from the performance of many analytics development publications it also becomes obvious that the apparently good output metrics of various workflows (given data favoring vysen above) are primarily due to the fact that these treatment data were generated with technical replicates, hence having errors that can be easily modeled and even defies the reasons often given for failure of statistics-based analytics.

The in-depth analysis in this whitepaper indicates (i) even considering a variety of them, performance metrics can be misleading, so the choice of dataset(s) is very important, and (ii) the concluding picture (boxplot) shown here indicates that the DEGs and Non-DEGs (true positives and false positives) unique to ReSurfX::vysen have high technical significance. This sets the stage for the upcoming whitepaper #2 on the application of these methods to complex, real-world data to evaluate these workflows for accuracy, ability to derive novel insights, and amenability to automate knowledge extraction.

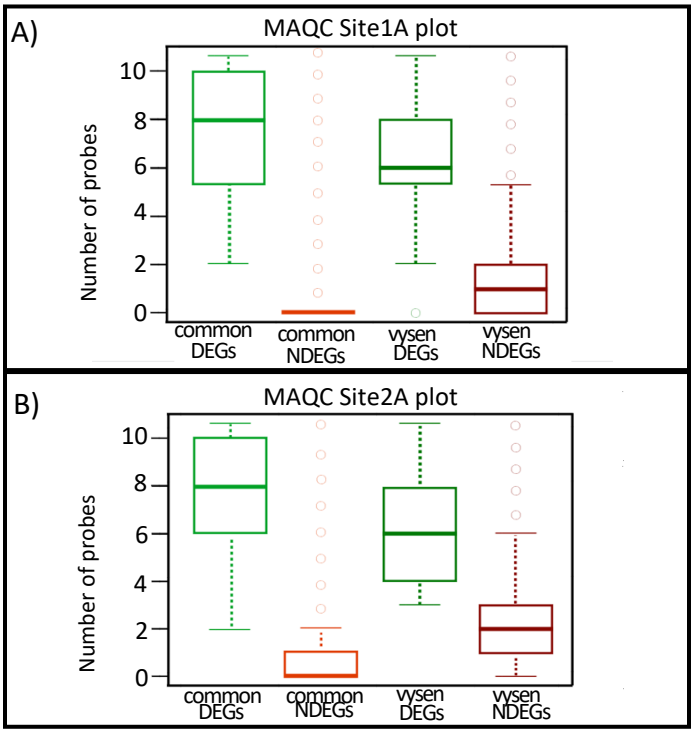
The choice of dataset used in deriving the performance metrics is an important consideration in choosing analytics suited for your data-analytics applications.

Fig. 12A, B, C, & D—The overlap of DEGs identified by vysen and other workflows used in this study



The Venn diagrams indicate the results from GeneChip are in for (i) Site1 and (ii) Site2. The equivalent results for RNASeq are in (iii) Site1 and (ii) Site2. The analysis of the differences like an example shown in Fig. 13 one showed that DEGs and Non-DEGs unique to vysen compared to the SWF should be very informative in making decisions on choice of analytics.

Fig. 13A & B—Number of independent measures of the genes used by ReSurfX::vysen in making a decision



Boxplots indicate GeneChip data of ReSurfX::vysen and SWF1A – for DEGs and Non-DEGs common to the workflow or unique to one of them. Note that the DEGs unique to vysen are false positives of SWF1 and vice-versa. The number of independent measures of a gene (probes) out of 11 independent measures are indicated in the y-axis as a boxplot for all the genes in each category.

We welcome discussions of these results and our product further by contacting https://resurfx.com/about/contact/

References

1. Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2. **Adaptive Hypersurface Technology (AHT) is motivated by the concept in** Gopalan S (2004) ResurfP: a response surface aided parametric test for identifying differentials in GeneChip based oligonucleotide array experiments. Genome Biology 5:P14

3. Shi L, et al. (2006) The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24:1151–1161.

4. Consortium S.M.-I. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat. Biotechnol. 32:903–914.

5. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, et al. (2010) Ontology-Based MetaAnalysis of Global Collections of High-Throughput Public Data. PLoS ONE 5(9): e13066. **Used in the product NextBio now part of BaseSpace product suite of Illumina.**

6. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). "affy—analysis of Affymetrix GeneChip data at the probe level." Bioinformatics, 20(3), 307–315. ISSN 1367-4803.

7. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, 43(7), e47.

8. Pollard KS, Dudoit S, van der Laan MJ (2005). Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer.

9 HISAT: a fast spliced aligner with low memory requirements. Nature Methods 2015.

10. Liao Y, Smyth GK and Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 30 (7):923-30, 2014.

11. Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140.

12. AHT normalization for RNA-seq data is conceptually similar to negative bionomial based normalization in subread package.

13. McCarthy, J. D, Chen, Yunshun, Smyth, K. G (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." Nucleic Acids Research, 40(10), 4288-4297.