Improving Outcomes Through Enhanced Data Analytics and Artificial Intelligence

Part 2: Gaining Novel Insights from Complex Real-World Data—ReSurfX::vysen vs Widely Used Analytics

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, organizations should adopt this digital revolution in order to be successful. We elaborate on our focus on comparative evaluation of a variety of data and workflow components to achieve superior accuracy, novel insights, automated knowledge extraction, and ROI improvements through intelligent allocation of experts, expertise, and resources.

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), machine learning, the cure for limitations of statistics and enterprise productivity.

In this second part, we test Adaptive Hypersurface Technology (AHT)[1] leveraged in our product ReSurfX::vysen to popular alternative analytical tools and knowledge bases built from them. From the resulting evaluation, we demonstrate that ReSurfX::vysen not only performs robustly with lesser data, but uniquely allows for deeper (causal) insights on components and interactions resulting in higher-level (inferential) knowledge. We want to highlight that the datasets from these two technologies are used to showcase the power of AHT where these two raw data have two different properties of multiple inputs use to measure an output.

This is a follow-up on our previous feature (**whitepaper #1**) which provided extensive information on the different data-analytics workflows used and the metrics evaluated. That previous feature extensively analyzed ReSurfX::vysen and a variety of other data analytics workflows with various metrics and showed that (i) user skepticism often arises because those (theoretically sound) metrics by themselves can be misleading, and (ii) choice of data plays a key role in making the right choice in choosing between alternatives.

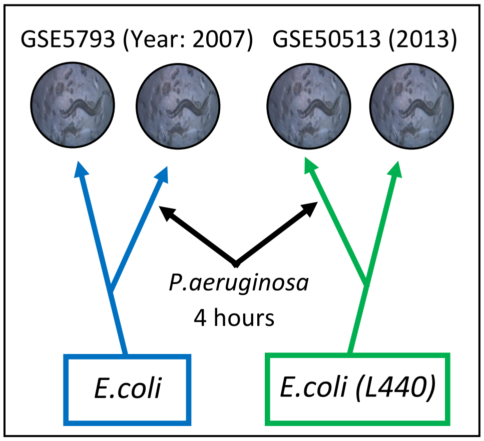
Here, we elaborate on these findings using real-world data to make critical points, and then set the stage for the use of these analytics and comparison of the outcome in larger-scale datasets. Some key insights include: (i) many analytics workflow breakdown when exposed to a wider variety of error properties (unknown error properties, as we called them in Part 1), (ii) the need for explicit tests of robustness and clear ways to evaluate accuracy and knowledge content, (iii) testing amenability of analytics to gain superior and in-depth knowledge of how an outcome is orchestrated, and (iv) identified a novel target for lipid-modification.

We highlight the features of the different workflows using examples from two classes of real-world data: (i) data from genome-wide analysis from the Ausubel laboratory (a previous ReSurfX user), and (ii) part of toxicogenomics dataset generated by MAQC/SEQC data[2] from the liver of rats treated with different chemicals. In the first case, we highlight essential information including identification of a key target proven to have phenotypic impact by genetic analysis though we analyzed all the data from that laboratory. In the second case, we use the first three chemicals from that MAQC/SEQC study. We end this whitepaper highlighting some properties of a much larger toxicogenomics study that will be further elaborated in Part 3 of this whitepaper series.

Real-world datasets used for detailed analysis presented in this study

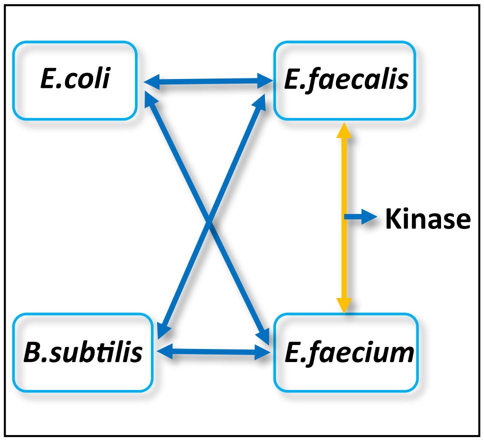
Here we present results from **DATASET1:** Affymetrix GeneChip data comparisons based on *C. elegans* fed with *E.coli* as food source (as control) and when exposed to human pathogen *P. aeruginosa* (PA14) for 4 hrs[3] (GSE5793) and slightly variant control *E.coli* bacteria[4] at a different point in time from another laboratory (GSE50513). The study design is depicted in Fig.1. There were subtle differences in the materials used but otherwise the studies were largely very similar; **DATASET2:** Affymetrix GeneChip data from *C. elegans* fed with closely related pathogens *E. faecium* and *E. faecalis* and GeneChip based gene expression data was collected from these and two different bacteria not considered pathogenic (*E.coli* and *B.subtilis*)[5]. The study design for this dataset is shown

Fig. 1—Study design DATASET1 of two control to treatment style experiments generated at different time points



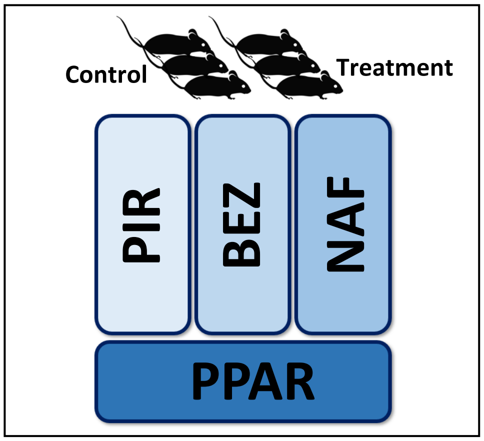
The differences between the two studies are (i) done at different time point and laboratories, and (ii) the control *E.coli* had some difference. Data were generated using GeneChip. [*C.elegans* picture courtesy: Genetics Society of America.]

Fig. 2—Study design for DATASET2 – comparing response of two closely related pathogens



The differential expression of a kinase between *E.faecalis* and *E.faecium* was identified by vysen and not by SWF1 and SWF2. This target was genetically proven to have significant difference phenotypic response between the two closely related pathogens in this model system.

Fig. 3—Study design of DATASET3 gene expression in rats fed with different lipid modulating chemicals



The known major pathway involved in the mode of action of these three chemicals is PPAR signaling pathway. These are the first three chemicals from the MAQC/SEQC toxicogenomics study[1]. Data generated from same tissue was available for microarray (GeneChip) and RNA-seq (using Illumina platforms).

in Fig. 2. **DATASET3:** Gene expression data from same tissue samples: (i) RNASeq data collected using Illumina platform and (ii) Affymetrix GeneChip data a subset of MAQC/SEQC study on rat liver exposed to different chemicals[1] – here we use first three chemicals of that study that have the same pathway as known dominant mode of action which gives sufficient information. We end this whitepaper with exemplary results from a larger dataset from NTP on gene expression in rat organs after exposure to a range of chemicals, upon which we will elaborate in the next whitepaper. The study design for this dataset is given in Fig. 3.

Box. 2 presented at the end of this whitepaper has a condensed reminder about the different workflows extensively tested with various performance metrics in **whitepaper #1.**

Both SWF1 and SWF2 are not robust when applied to real-world data

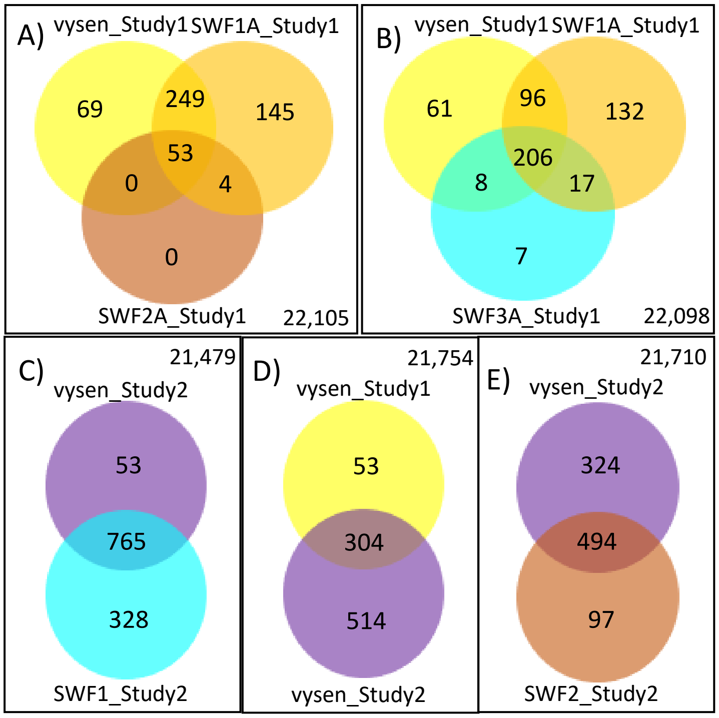
SWF2 breaks down on a simple study design (Fig. 1, Fig. 4A) with real world data. In this case we use data from Study1 – i.e., one of the two laboratories – Fig. 1. Clearly, the number of DEGs is considerably low for SWF2 compared to SWF1, SWF3, or vysen. That SWF2 result is indeed an error in the DATASET1, can be confirmed by the overlap between DATASET1 and DATASET2 (nearly identical experiment between two laboratories) indicated about 85% of DEGs from Study1 was identified in Study2, and about 35% of Study2 overlapped with Study1 (except for SWF1) (Fig. 4A – 4E).

Given SWF2A is a reasonably simplistic workflow; normalized and summarized GeneChip data were subject to t-test with p-value, false discovery rate (fdr – Benjamini and Hochberg approach) and ratio thresholding – this likely makes sense, as the data and error properties get more complex.

SWF3 identified far fewer DEGs than vysen or SWF1 – as had been observed in whitepaper1 with technical replicate data based comparison and with other datasets (except in case of failure of workflows SWF1 or SWF2). vysen identified about 80% of DEGs, by simple count, compared to SWF1of in both studies. Study of such non-overlapping DEGs (i.e., DEGs unique to SWF1 and vysen) in Fig. 13 of **whitepaper #1** indicated that vysen results had compelling meaning, both when considering DEGs uniquely identified by vysen or DEGs excluded from those identified by SWF1.

SWF2A also identified only 2 DEGs from DATASET3-BEZ due increased unpredictability of SWF2 over large amounts of data than other workflows discussed in this whitepaper series. Surprisingly, SWF1A also breaks down frequently despite employing relatively more sophisticated eBayes for error property dependent data value determination and linear modeling to specify the class

Fig. 4—Differential expressed genes found by different workflows for DATASET1 in Fig. 1



SWF1A, SWF2A, SWF3A represent workflows compared with vysen for analysis of gene expression data from GeneChip platform. SWF1R and SWF2R were workflows compared with vysen for analysis of RNA-seq based gene expression data. This data clearly demonstrates failure of SWF2A.

> AHT based ReSurfX::vysen analytics has better sensitivity, more knowledge content, and is more amenable to automation

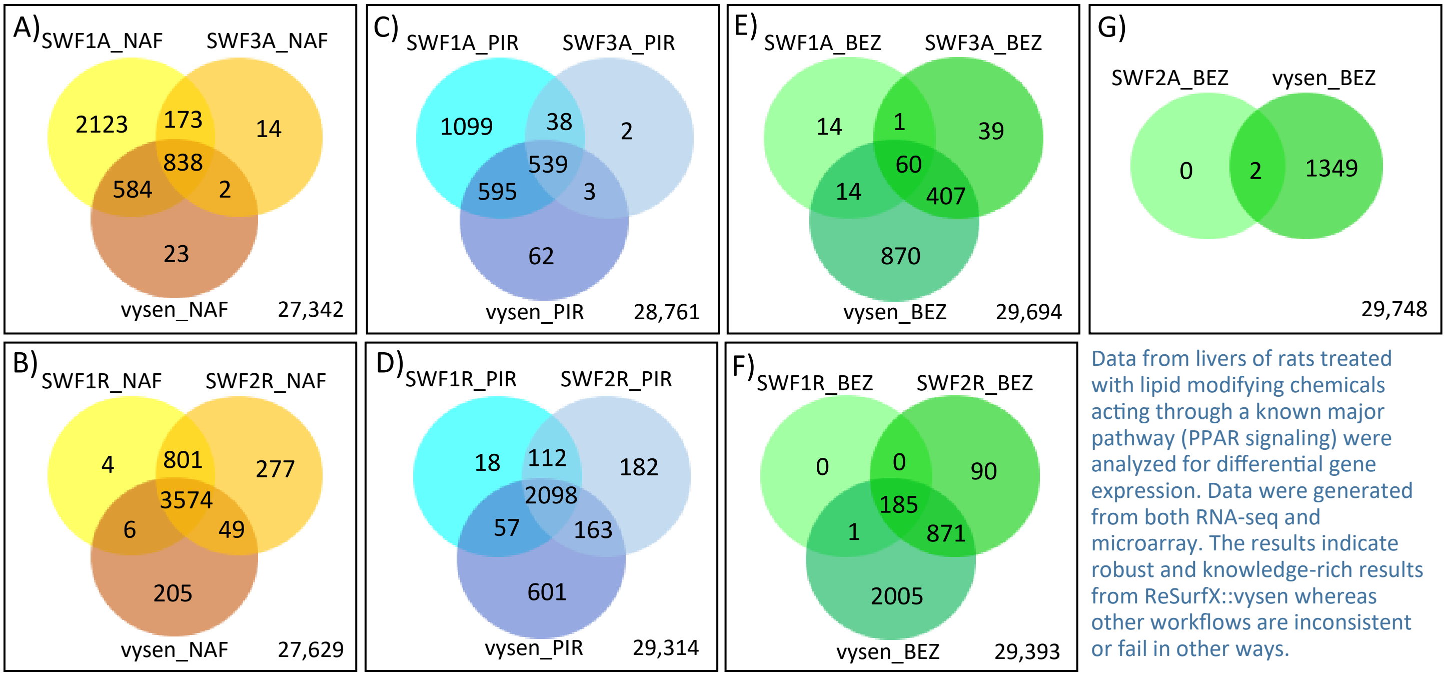
labels of the samples used in comparative analytics. This is exemplified here, as SWF1A identified very low number of DEGs in DATASET3-BEZ compared to SWF3A, and vysen (Fig. 5E). Interestingly, SWF1R for RNASeq, which also uses linear modeling and eBayes on the data from same tissue (DATASET3-BEZ), similarly identified a very low number of DEGs compared to SWF2R and vysen (Fig. 5F).

We indicate such consistent SWF1 failure over a large number of comparisons at the end of this whitepaper (Fig. 6 and Fig. 7). Just for the 27 chemicals from this rat toxicogenomic study[2] SWF1 yielded less than 5 DEGs in 7 of them – whereas ReSurfX::vysen and SWF3 identified considerably higher number of DEGs in those cases.

vysen analytics has better sensitivity and is more automation amenable

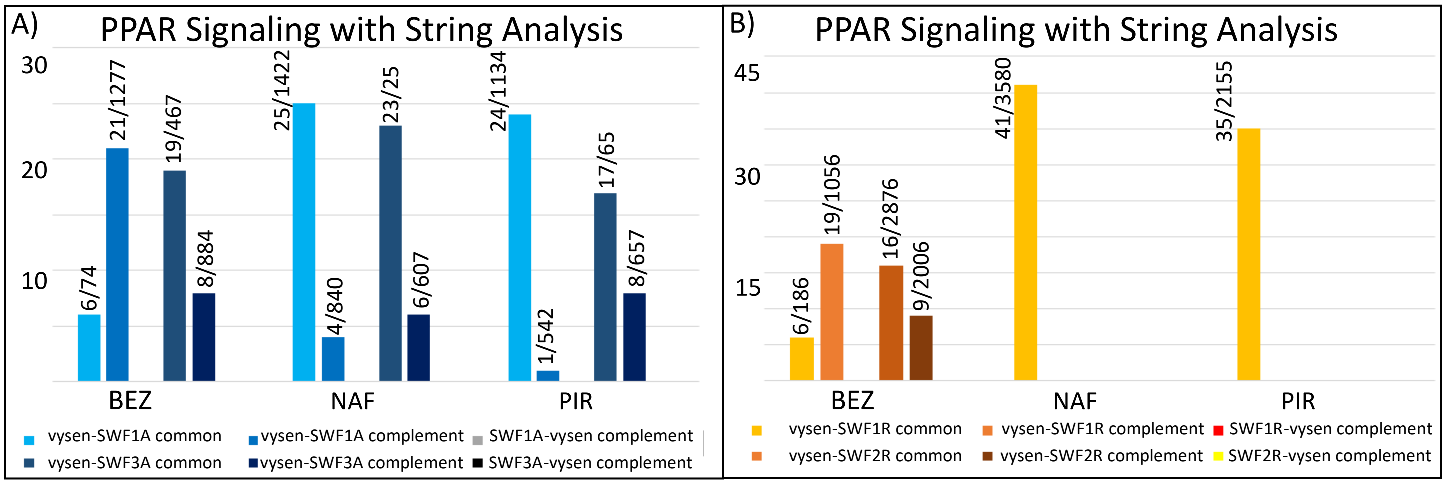
In the study design depicted in Fig. 2 that compares gene expression response of *C. elegans* feeding on two closely related human pathogens (*E. faecalis* and *E. faecium*) vysen (and not SWF1 or a workflow used in ref 5) identified *mpk-2* as an interesting target upon which to expend resources. Genetic elimination revealed that this kinase had an effect on the disease phenotype of *C. elegans* between the two pathogenic bacteria. Nearly all other DEGs identified by vysen as DEGs between these closely related pathogens were either genes duplicated on the chip or closed related members of gene families. Looking at each difference between the non-pathogenic control bacteria (*E.coli* and *B. subltilis*) and each of the two pathogens might have identified *mpk-2* as a pathogen associated kinase amongst lot of other DEGs. An automated workflow will likely exclude or rank that gene lower on priority when looking for differences between the two pathogens because of missing transitivity.

Fig.5—Gene expression analysis rats livers demonstrate failure of multiple widely used analysis workflows



Data from livers of rats treated with lipid modifying chemicals acting through a known major pathway (PPAR signaling) were analyzed for differential gene expression. Data were generated from both RNA-seq and microarray. The results indicate robust and knowledge-rich results from ReSurfX::vysen whereas other workflows are inconsistent or fail in other ways.

Fig. 6—The number of PPAR signaling genes identified in different comparative differentially expressed subset between different analytic workflows



PPAR signaling is the known major pathway for mode of action (MOA) of these three chemicals. Results shown are for GeneChip microarray (A) and from RNA-seq data (B). **The results from differentially expressed genes indicate robust performance of vysen and either failure or sub-standard performance of other workflows.** This analysis was done using the STRING database[7].

ReSurfX::vysen analysis yield robust, consistent and automation amenable results with and richer and novel insights while other workflows are unreliable

DATASET3[2] has the first three chemicals of rat toxicogenomics study of MAQC/SEQC consortium – nafenopin (NAF), pirinixic acid (PIR) and bezafibrate (BEZ). A flavor of large scale performance of these workflows on GeneChip data is shared at the end of this whitepaper.

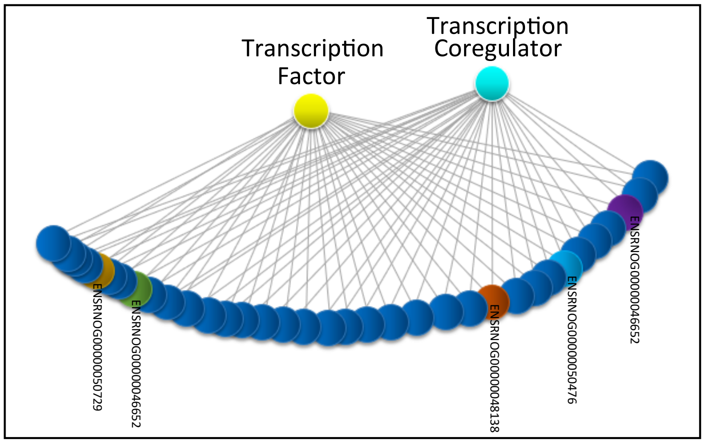
These three lipid-modifying chemicals are chosen for display in this whitepaper as they show the range of information we want to highlight, there were the first three from that study (avoiding cherry-pick) and the major known biological pathway these chemicals act are the same – i.e., **the PPAR signaling pathway**.

As shown in the Fig. 5A to Fig. 5F, both GeneChip data and RNA-seq data displayed a variety in extent of overlap as well as DEGs unique to one of the workflows. The results show that the widely used as well well as a commercial workflow are erratic in performance or get far fewer DEGs (even without considering the false positives present in them). The following are the most notable observations:

- Both SWF1A for GeneChip and SWF1R for RNA-seq failed in the analysis of chemical BEZ. Both these have common properties in the workflow (e.g., linear modeling, eBayes).
- SWF2A failed with DATASET3-BEZ (as it also did with the DATASET1).
- SWF3 consistently found far lower number of DEGs.
- In all cases, as is expected with SWF3 being less sensitive, the DEGs found by vysen alone in comparison to SWF3 always had genes representing PPAR pathway (Fig. 6). This PPAR pathway is known to be a major determinant of mode of action (MOA) of these three chemicals.
- DEGs found by vysen alone often had PPAR pathway related genes despite removing the common DEGs (hence expected to be enriched in the major pathway components) (Fig. 5 and Fig. 6).
- Despite SWF1A and SWF1R having 1000-2000 more DEGs than corresponding vysen, analyzed results from rats treated with nafenopin and pirinixic acid, respectively, did not contain any PPAR pathway related component (Fig. 5 and Fig. 6).
- The large number of extra DEGs found by vysen analysis even in comparison with SWF2R which did find a large number of DEGs that overlapped with vysen had:
    - Very rich in network connectivity – suggesting good amount of knowledge of these genes acting together in other biological studies.

Importantly, this set of DEGs identified by vysen alone revealed an obvious network section connecting a large number of terminal nodes to a transcription factor and a transcriptional coregulator (Fig. 7). Interestingly nearly all the nodes (genes) connected to the common transcription factor, coregulator pair belonged to a single gene family. This will be target molecule ripe for further analysis in a discovery workflow. We did not do additional network decomposition analysis of the remaining genes of this class (DEGS identified by vysen alone) from RNA-seq dataset.

Fig. 7—A novel pathway of significant interest in differentially expressed gene identified by ReSurfX::vysen alone



A previously unidentified relationship in the context of response to lipid modifying chemical bezafibrate was identified by ReSurfX::vysen from RNA-seq data. **Interestingly, all the nodes in blue that the transcription factor and a transcription coregulatory connect to belong to one gene family known to have effect on lipids in higher organisms.** This analysis was done using the STRING database[7].

Thus AHT-based ReSurfX::vysen identifies information rich information as well as leave out (eliminate) information poor (misleading) information. This should be remarkably empowering to enterprise ROI.

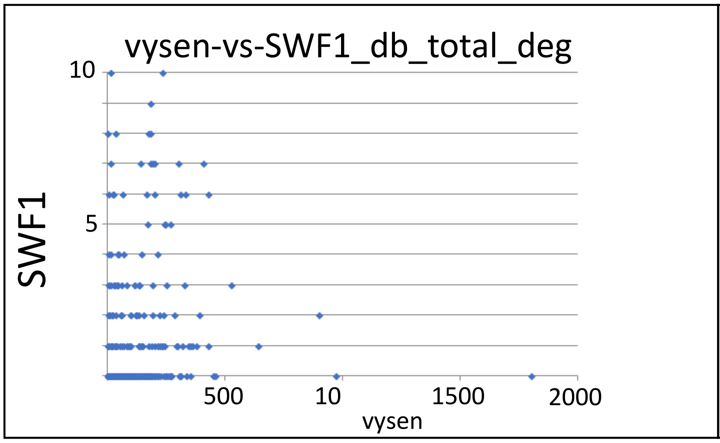
Large scale analysis revealed widespread failure of a widely used GeneChip workflow and significant under-detection of DEGs by a commercial workflow

We are assembling are large-scale drug response dataset that can be queried with vysen and other analysis in the ReSurfX::vysen product. Initially we have chosen data that would be considered to have smaller number of replicates far below statistical power but typical of discovery workflows. This analysis reveals large scale failure of a widely used workflow (Fig. 8) as can be seen by finding ZERO DEGs in a very large number of comparisons.

This analysis also revealed that the SWF3 workflow[6] used in the commercial product, previously NextBio – now part of BaseSpace suite of analytics products from Illumina identifies less than 50% DEGs than when analyzed using ReSurfX::vysen (Fig. 9). **This indicates that even if one can find dominant pathways representing mode of action of drug molecules, the results in that product using SWF3 workflow will be insufficient to detect how the phenotypic outcome or the dominant pathways are modulated.**

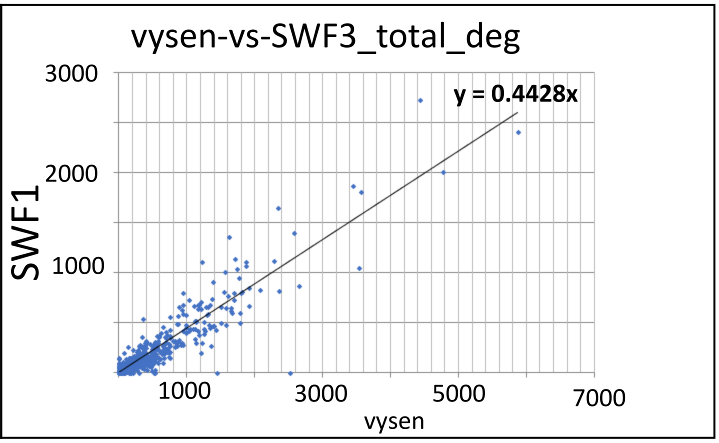
The database briefly outlined above is available for query together with the novel analytics based on AHT is provided in ReSurfX::vysen SaaS product, and will be the topic of more detailed discussion in the upcoming **whitepaper #3**.

Fig. 8—Widespread failure of SWF1 while ReSurfX::vysen is robust in a mini-database sized analysis



The results are derived from a part of (650 comparisons) of a drug response database being created and available in ReSurfX::vysen product. The y-axis indicates the instances where SWF1 workflow identified 10 DEGs or less. SWF1 had 0 DEGs in 231 comparisons. **ReSurfX::vysen had 0 DEGs in 4 comparisons, hence it was more robust and informative.**

Fig. 9—Consistently low number of DEGs are identified by a commercial workflow while ReSurfX::vysen is robust in a mini-database sized analysis

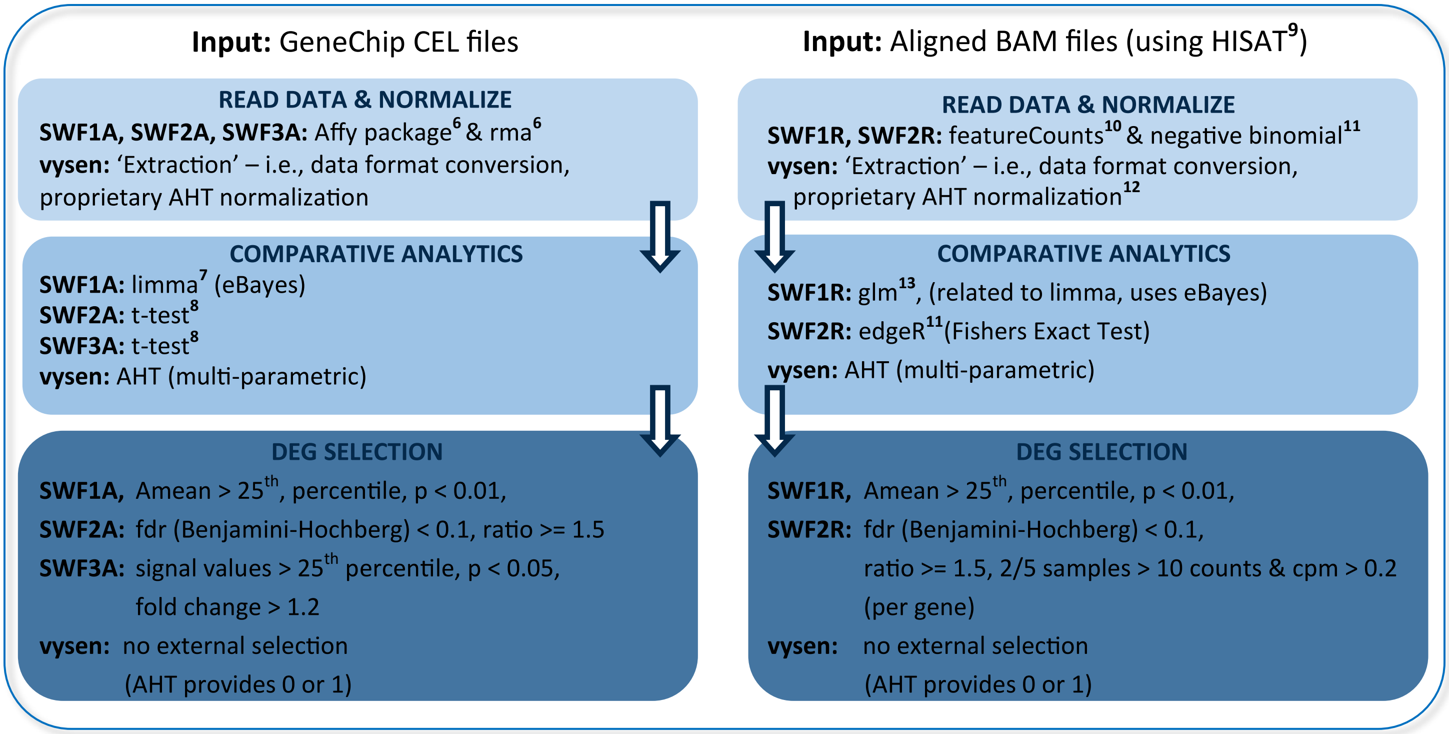


The results are derived from a part of (650 comparisons) of a drug response database being created and available in ReSurfX::vysen product. The number of DEGs identified by vysen (x-axis) plotted against the number of DEGs identified by a commercial workflow (SWF3) is less than 50% on average. This, with data from Fig. 6 demonstrates superior performance of ReSurfX::vysen to study MOA.

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

*Affymetrix, GeneChip, Illumina, ReSurfX, Adaptive Hypersurface Technology, AHT, BaseSpace are trademarks and/or copyrights belonging to one of Thermo, Illumina Corp, and ReSurfX, Inc.

Box 2



Input: GeneChip CEL files

READ DATA & NORMALIZE

SWF1A, SWF2A, SWF3A: Affy package[6] & rma[6]

vysen: 'Extraction' – i.e., data format conversion, proprietary AHT normalization

COMPARATIVE ANALYTICS

SWF1A: limma[7] (eBayes)

SWF2A: t-test[8]

SWF3A: t-test[8]

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)

Input: Aligned BAM files (using HISAT[9])

READ DATA & NORMALIZE

SWF1R, SWF2R: featureCounts[10] & negative binomial[11]

vysen: 'Extraction' – i.e., data format conversion, proprietary AHT normalization[12]

COMPARATIVE ANALYTICS

SWF1R: glm[13], (related to limma, uses eBayes)

SWF2R: edgeR[11](Fishers Exact Test)

vysen: AHT (multi-parametric)

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)

References

1. **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

2. Gong, B. *et al.* Transcriptomic profiling of rat liver samples in a comprehensive study design by RNA-Seq. *Sci. Data* 1:140021 doi: 10.1038/sdata.2014.21 (2014).

3. Troemel ER, Chu SW, Reinke V, Lee SS, Ausubel FM, Kim DH (2006) p38 MAPK Regulates Expression of Immune Response Genes and Contributes to Longevity in C. elegans. PLoS Genet 2(11): e183.

4. Estes, Kathleen A. et al. bZIP Transcription Factor *zip-2* Mediates an Early Response to *Pseudomonas Aeruginosa* Infection in *Caenorhabditis Elegans* . *Proceedings of the National Academy of Sciences of the United States of America* 107.5 (2010): 2153–2158.

5. Yuen, Grace J., and Frederick M. Ausubel. "Both Live and Dead *Enterococci*Activate *Caenorhabditis Elegans* Host Defense via Immune and Stress Pathways." *Virulence* 9.1 (2018): 683–699.

6. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, et al. (2010) Ontology-Based Meta-Analysis 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.**

7. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017 45:D362-68