A Workbench using Evolutionary Genetic Algorithms for analyzing association in TCGA Data

Gilmore, A., Alderdice, M., Savage, K., O'Reilly, P. G., Roddy, A., Dunne, P., Lawler, M., McDade, S., Waugh, D., & McArt, D. (2019). A Workbench using Evolutionary Genetic Algorithms for analyzing association in TCGA Data. *Cancer Research*, *79*(8), 2072. https://doi.org/10.1158/0008-5472.CAN-18-1976

Published in:
Cancer Research

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Cancer Research - Resource report

ACE: A Workbench using Evolutionary Genetic Algorithms for analyzing association in TCGA Data

Alan R Gilmore [1#], Matthew Alderdice [1], Kienan I Savage [1], Paul G O'Reilly [1], Aideen C Roddy [1], Philip D Dunne [1], Mark Lawler[1], Simon S McDade [1], David J Waugh [1^]and Darragh G McArt [1^#]

[1]Centre for Cancer Research and Cell Biology, Queen's University Belfast, 97 Lisburn Road, Belfast, Northern Ireland, BT9 7AE, United Kingdom

To whom correspondence can be addressed.

^ Joint Senior authorship

Corresponding authors

Darragh G McArt

Email - d.mcart@qub.ac.uk
Phone - +44 (0)28 9097 2629
Mailing Address - Centre for Cancer Research and Cell Biology, Queen's University Belfast, 97 Lisburn Road, Belfast, Northern Ireland, BT9 7AE, United Kingdom

Alan R Gilmore
Email - a.gilmore@qub.ac.uk,
Phone - +44 (0)28 9097 2629
Mailing Address - Centre for Cancer Research and Cell Biology, Queen's University Belfast, 97 Lisburn Road, Belfast, Northern Ireland, BT9 7AE, United Kingdom

Disclosure of Potential Conflicts of Interest

The authors declare no conflicts of interest

Running Title: Evolutionary algorithm for biomarker discovery in cancer

Keywords: Evolutionary algorithm, visualization, biomarker correlation

Grant Support
A. C. Roddy and D. G. McArt received CRUK : C11512/A20877. M. Alderdice received MRC:

MR/S003789/1

Abstract

Modern methods of acquiring molecular data have improved rapidly in recent years, making it easier for researchers to collect large volumes of information. However, this has increased the challenge of recognizing interesting patterns within the data. Atlas Correlation Explorer (ACE) is a user-friendly workbench for seeking associations between attributes in the cancer genome atlas (TCGA) database. It allows any combination of clinical and genomic data streams to be searched using an evolutionary algorithm approach. To showcase ACE, we assessed which RNA-sequencing transcripts were associated with estrogen receptor (ESR1) in the TCGA breast cancer cohort. The analysis revealed already well-established associations with XBP1 and FOXA1, but also identified a strong association with CT62, a potential immunotherapeutic target with few previous associations with breast cancer. In conclusion, ACE can produce results for very large searches in a short time and will serve as an increasingly useful tool for biomarker discovery in the big data era.

Significance: ACE uses an evolutionary algorithm approach to perform large searches for associations between any combination of data in the TCGA database.

Introduction

The steady improvement of sequencing technology is leading to a proliferation of large collections of molecular data being used in the pursuit of precision medicine. In turn, analyzing or searching these collections in conventional ways may take an impractical amount of time, making it inevitable that many anomalies and insights hidden in this data will be overlooked(1).

In an effort to address these challenges we designed a software environment to allow easier exploration of such data collections for new biological insights. The framework centers on an evolution-based data-mining engine. The aim is to highlight biological relationships and insights that a human expert would be unlikely to observe. The initial implementation is based on *TCGA* ("The Cancer Genome Atlas") data (2) and was therefore coined *ACE* ("Atlas Correlation Explorer").

The core technology of *ACE* is an engine which uses a genetic approach modeled on evolution to carry out its search. Such approaches are known to be effective in finding near-optimal solutions to computationally-intensive problems in much shorter times than other methods, and they have been applied in many differing fields (3-5).This engine creates a large pool of software "organisms", each of which looks for correlation between specific molecular measurements in *TCGA* data. As with real-world evolution it is indiscriminate in the initial selection, so for the vast majority of organisms no correlation is found. However, when any correlation is found, that organism will thrive in the evolutionary pool, and will be refined through the equivalent of mating and mutation. The engine creates large numbers of such organisms, and once evolution has been allowed to continue for some time, the best organisms will have identified some significant connections. This engine was originally developed to address pattern analysis applications in financial markets, where the problems were too complex to be addressed by "brute force" methods, and a genetic approach was

found to allow convergence to near-optimal solutions in practical timescales. A more detailed description of the ACE algorithm can be found in supplementary file appendices.

Materials and Methods

Architecture and workflow

The data used by the engine was derived from the Broad Firehose TCGA pipelines (6), but has been converted into compact binary files that can be loaded directly into memory for very fast access. A more detailed description of the binary files can be found in the supplementary file appendices. Data associated with successful “organisms” are cached for re-use. ACE is implemented as a *Windows* desktop application with requirements of *Windows 7* or later, but can also be used on OS/X or Linux using Windows VM technology. ACE was developed using C# and *Microsoft Visual Studio*. Approximately two gigabytes of space are required for the converted data of each cancer type. In using *ACE*, a researcher can select subsets of the *TCGA* data, and *ACE* will search the values in those subsets and pick out those that show strong correlation. When the chosen subsets are large, an exhaustive search for the best correlations would be impractical, but the evolutionary nature of *ACE* allows for good associations to be found in reasonably short times. This engine is flexible in application, allowing systems using it to define their own “universe”, to define the type of organisms that will exist within it, and to define mutation and mating algorithms relevant to that universe. The “survival of the fittest” principle is used, but each system using the engine can define a scoring system for organisms, letting only the highest-scoring individuals survive to produce offspring. This allows a system implemented using the engine complete flexibility in setting goals for the evolution.

Data and Usage

ACE has two main screen views, and a user can toggle between them. The Data Selection View allows browsing of the TCGA data, and selection of portions of it to be analysed by ACE. A biomarker or feature of interest should be selected as a source measure. Target

measures are the features from which association strength will be made with the source measure (see figure 1 and video 1). The Genetic Search View is used to run an evolution based on the selected data, and to view a ‘leader board’, showing the best correlations found by ACE between items of the selected data. Each correlation within the leader board can be selected and visualized (see figure 1 and video 1).

An import and conversion phase is needed to prepare the TCGA data for each cancer type for use with ACE. Currently the following cancer types are supported: Bladder urothelial carcinoma (BLCA); Breast invasive carcinoma (BRCA); Colon adenocarcinoma (COAD); Glioblastoma multiforme (GBM); Head and Neck squamous cell carcinoma (HNSC); Lung adenocarcinoma (LUAD); and Prostate adenocarcinoma (PRAD) (see video). Additional cancer types will be included in future releases and a comprehensive manual can be found in the supplementary materials and methods.

The addition of miRNAseq and methylation data for the available cancer types vastly increases the potential pool of data points for correlation analysis. This represents an important advantage ACE has over portals which also assess correlations within and across data types (7). Another important feature within ACE is the inclusion of multiple data processing outputs. It is well recognized that different methods of data processing can have profound effects upon the downstream interpretation(8). There is currently a lack of consensus on which normalization and transformation pipelines result in the most robust outputs from high-throughput array and sequencing technologies. For array technologies, collapsing multiple probes or using median or most variable values can reduce computational requirements, improve signal-to-noise ratios and facilitate cross platform analysis, however, these data reduction processes can result in loss of information(9). The same principle applies for sequencing technologies where different processing pipelines can result in different observations from the same data. On this basis, ACE has incorporated different levels of normalization and transformation across all data types where possible (see video). This permits researchers to assess diverse pipelines that may yield variable but

potentially clinically relevant observations from the same data. Unlike the correlative analyses that are performed in similar TCGA portals which output hundreds and potentially thousands of correlations, ACE outputs a more succinct leader board of associations(n=36) from a vast pool of data. Although this succinct output has inherent data loss caveats, it allows researchers to focus upon a more concise list of correlated biomarkers to take forward for hypothesis generation and subsequent investigation. Assessing commonality across multiple runs would further help bolster the robustness of the observed correlations.

Results

To showcase ACE, we have used the well-defined breast cancer biomarker Estrogen Receptor (*ESR1)*, using mRNAseq_RPKM_log2 expression as the source measure and all other mRNAseq_RPKM_log2 features as the target measures within the TCGA breast cancer dataset(10). The top target measure from this analysis is *Chromosome 6 Open Reading Frame 97*(*C6orf97)*also known as*(Coiled-Coil Domain Containing 170)CCDC170*, which is a transcription factor already known to be critical for the expression of *ESR1 in* breast cancer(11). Other well-known hits include Carbonic anhydrase 12 (*CA12), Forkhead box protein A1 (FOXA1) and* X-box binding protein 1 (*XBP1)* expression which have been extensively linked to estrogen receptor positive breast cancer (12-14).However, interestingly, *Cancer/Testis Antigen 62 (CT62)*, was the third strongest association from the analysis. Although dysregulated expression of cancer testis antigens (CTAs) has been observed in breast cancer before, there have been very few previous associations made with*CT62* and estrogen driven breast cancer(15).

Having identified *CT62* as a feature of interest associated with *ESR1*, we next assigned *CT62* as the source measure to assess the strongest correlations across all other features within the mRNAseq_RPKM_log2 dataset. Importantly, *ESR1, XBP1* and *FOXA1* featured in the analysis leader board which further indicated that CT62 is associated with estrogen-driven biology (Supplementary figure 1). Next we exploited ACEs capability to extract

features across the multiple molecular and clinical data types available within the BRCA dataset. Using the reverse phase protein array (rppa) pipeline in ACE we observed that ESR1 protein expression was the most correlated feature with *CT62* followed by the estrogen associated protein GATA3 (Supplementary figure 2) (16, 17). This observation indicates that *CT62* is associated with estrogen driven biology not just at the mRNA level but also the protein level.

The expression of cancer/testis antigens is known to be regulated via epigenetic mechanisms, therefore, we interrogated the mean methylation and the miRseq_RPKM_log2 pipelines to extract features correlated with *CT62* mRNA expression. Interestingly, *FOXA1* methylation was observed to be inversely associated with *CT62* mRNA expression (Supplementary figure 3). *ESR1* expression is known to be regulated by *FOXA1*, and hypermethylation of *FOXA1* has been shown to be associated with ER-negative tumours, therefore, mechanistic investigation should be employed to assess whether *FOXA1* methylation also regulates *CT62* expression (18). By interrogating the miRNA profiles we observed *MIR190b* to be positively correlated with *CT62* gene expression (Supplementary figure 4). It has previously been demonstrated that *MIR190b* is the most differentially expressed microRNA between ER-positive and ER-negative breast cancers (19). Finally, using ACE we examined which clinical parameters were correlated with *CT62*. We observed that ER status was the strongest correlation, with increased *CT62* being positively correlated with ER-positive tumours (Supplementary figure 5). See supplementary data file for all exported ACE outputs.

Discussion

The aim of this analysis was to demonstrate that ACE can highlight both well-established and unknown features associated with the Estrogen Receptor positive breast cancer. Our ACE analyses within the mRNA, methylation, microRNA, protein and clinical pipelines of the TCGA BRCA dataset indicates that CT62 is unequivocally associated with estrogen-driven

breast cancer. As the success of immune checkpoint blockade in breast cancer has been modest, a greater understanding of immune oncology targets in breast cancer is required. CTAs such as CT62 represent promising immunotherapeutic targets and biomarkers for breast cancer patients. The recruitment process of a phase I clinical trial (NCT03110445) for a CTA vaccine in breast cancer patients is imminent (20, 21).

In conclusion, this simple study design demonstrates that the user-friendly workbench ACE can successfully highlight established biological associations within the breast cancer paradigm but also has the potential to highlight poorly reported associations that may further our biological understanding and clinical management of the disease. We hypothesize that in the big data era, the heuristic search approach adopted by ACEs genetic engine will become increasingly useful for biomarker discovery.

Word count – 1695

Acknowledgments

The authors would like to thank the bioinformatics department and the informatics core unit at Queen's University Belfast (QUB) for testing the framework and the Centre for Cancer Research and Cell Biology at QUB for informed feedback and discussions.

References

1. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods 2016;111:21-31.

2. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113-20.

3. Notredame C, O'Brien EA, Higgins DG. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res 1997;25:4570-80.

4. Aerts S, Van Loo P, Moreau Y, De Moor B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics 2004;20:1974-6.

5. Tumuluru JS, McCulloch R. Application of hybrid genetic algorithm routine in optimizing food and bioengineering processes. Foods 2016;5:10.3390/foods5040076.

6. Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run [homepage on the Internet]. Broad Institute of MIT and Harvard: Broad Institute TCGA Genome Data Analysis Center. 2016. Available from: https://doi.org/10.7908/C11G0KM9.

7. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401-4.

8. Yang C, Wu PY, Tong L, Phan JH, Wang MD. The impact of RNA-seq aligners on gene expression estimation. ACM BCB 2015;2015:462-71.

9. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for aggregating gene expression data: The collapseRows R function. BMC Bioinformatics 2011;12:322,2105-12-322.

10. Sommer S, Fuqua SA. Estrogen receptor and breast cancer. Semin Cancer Biol 2001;11:339-52.

11. Yamamoto-Ibusuki M, Yamamoto Y, Fujiwara S, Sueta A, Yamamoto S, Hayashi M, et al. C6ORF97-ESR1 breast cancer susceptibility locus: Influence on progression and survival in breast cancer patients. Eur J Hum Genet 2015;23:949-56.

12. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet 2011;43:27-33.

13. Barnett DH, Sheng S, Charn TH, Waheed A, Sly WS, Lin CY, et al. Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. Cancer Res 2008;68:3505-15.

14. Sengupta S, Sharma CG, Jordan VC. Estrogen regulation of X-box binding protein-1 and its role in estrogen induced growth of breast and endometrial cancer cells. Horm Mol Biol Clin Investig 2010;2:235-43.

15. Jonsson P, Coarfa C, Mesmar F, Raz T, Rajapakshe K, Thompson JF, et al. Single-molecule sequencing reveals estrogen-regulated clinically relevant lncRNAs in breast cancer. Mol Endocrinol 2015;29:1634-45.

16. Hoch RV, Thompson DA, Baker RJ, Weigel RJ. GATA-3 is expressed in association with estrogen receptor in breast cancer. Int J Cancer 1999;84:122-8.

17. Theodorou V, Stark R, Menon S, Carroll JS. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res 2013;23:12-22.

18. Espinal AC, Buas MF, Wang D, Cheng DT, Sucheston-Campbell L, Hu Q, et al. FOXA1 hypermethylation: Link between parity and ER-negative breast cancer in african american women? Breast Cancer Res Treat 2017;166:559-68.

19. Cizeron-Clairac G, Lallemand F, Vacher S, Lidereau R, Bieche I, Callens C. MiR-190b, the highest up-regulated miRNA in ERalpha-positive compared to ERalpha-negative breast tumors, a new biomarker in breast cancers? BMC Cancer 2015;15:499,015-1505-5.

20. Vonderheide RH, Domchek SM, Clark AS. Immunotherapy for breast cancer: What are we missing? Clin Cancer Res 2017;23:2640-6.

21. Gjerstorff MF, Andersen MH, Ditzel HJ. Oncogenic cancer/testis antigens: Prime candidates for immunotherapy. Oncotarget 2015;6:15772-87.

Figure Legends

Figure 1.

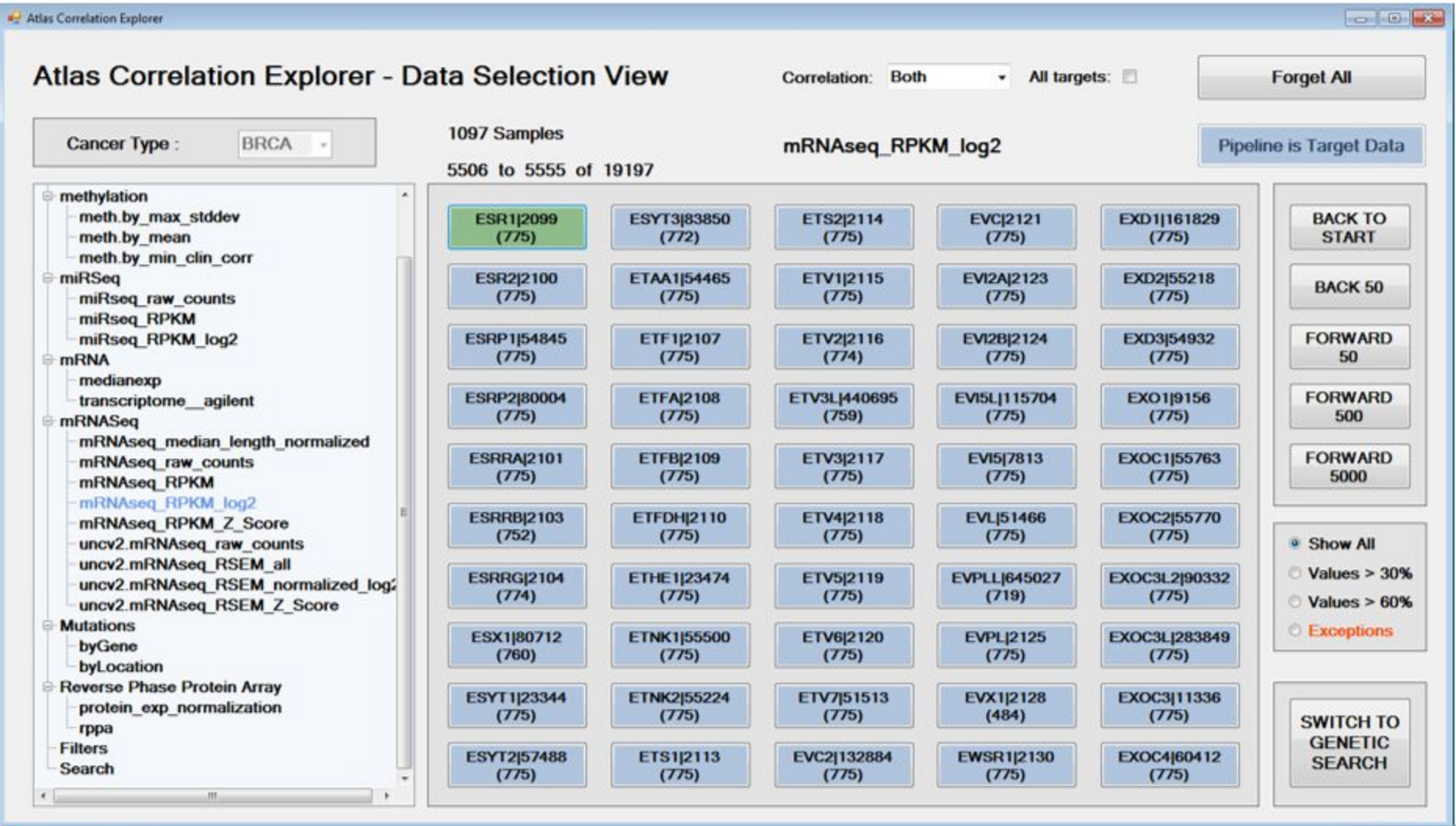
Top: This illustrates the ‘Data Selection View’. In this case the cancer type being studied is Breast Cancer (BRCA), the selected source measure is ESR1 from the mRNAseq_RPKM_log2 pipeline, while the target measures are all other features from the mRNAseq_RPKM_log2 pipeline. The right hand side shows the number of organisms generated and the heat map depicts the coverage achieved.

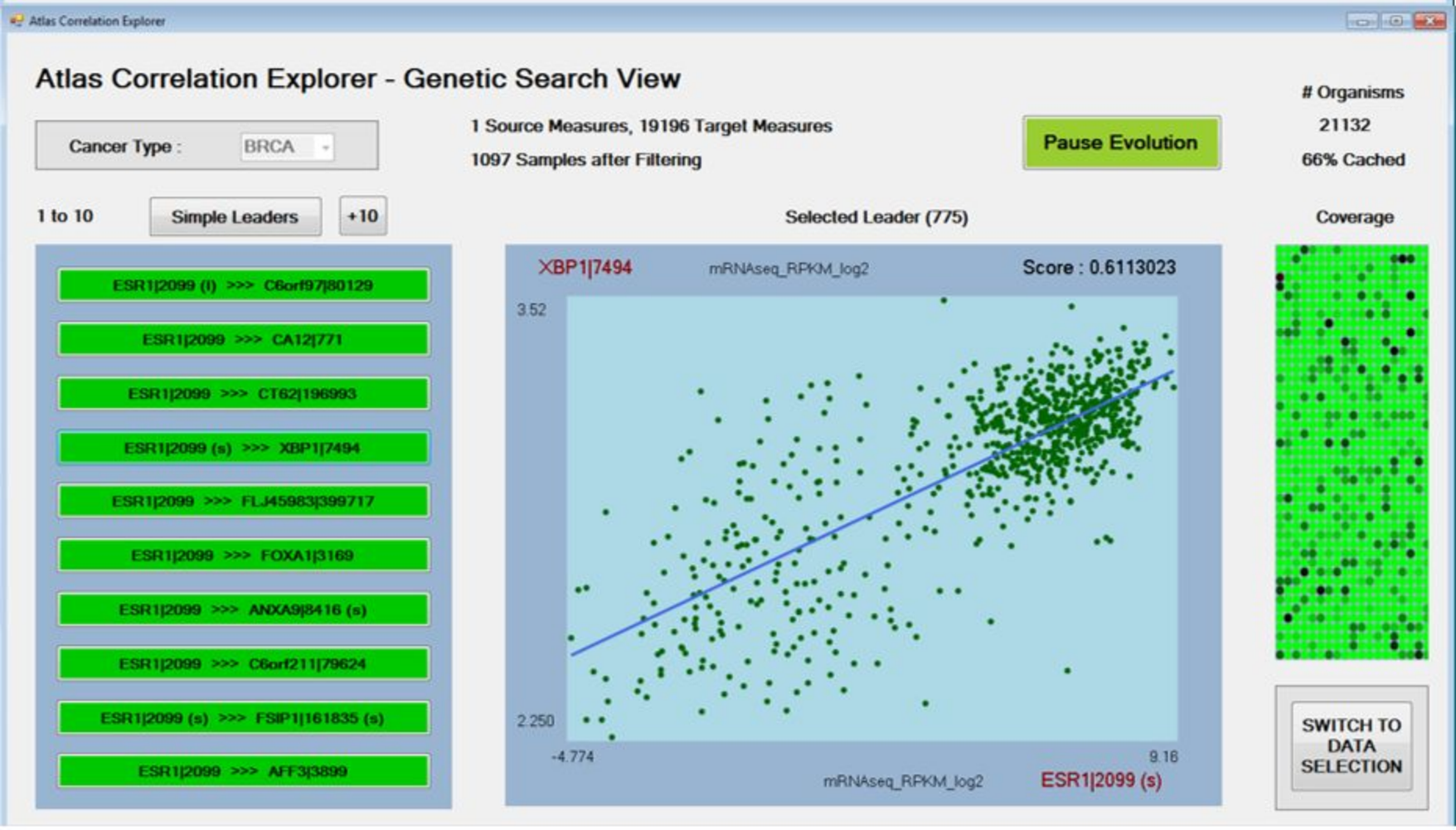
Bottom: This demonstrates the ‘Genetic Search View’ being used to search for correlation between expression, methylation, protein and mutational data with ESR1 mRNAseq_RPKM_log2 (s) as the marker of interest. The list generated dynamically on the left hand side represents the best correlations found so far, and the main panel shows strong association depicting top hits XBP1. The (s) indicates that a correlation has been found using the SQUARE ROOT of the given values. As part of the evolution, this is tried in addition to LOG (l) and ARCSIN (a) as investigation has shown that these sometimes provide a better fit than the raw values.

Footnotes

Note: Supportive information and use case data surrounding this article are available through the deposit on GitHub (https://github.com/AlanRGilmore/ACE)

Figure 1


Atlas Correlation Explorer
Atlas Correlation Explorer - Data Selection View
Correlation: Both
All targets:
Forget All
Cancer Type : BRCA
1097 Samples
5506 to 5555 of 19197
mRNAseq_RPKM_log2
Pipeline is Target Data
methylation
meth.by_max_stddev
meth.by_mean
meth.by_min_clin_corr
miRSeq
miRseq_raw_counts
miRseq_RPKM
miRseq_RPKM_log2
mRNA
medianexp
transcriptome__agilent
mRNASeq
mRNAseq_median_length_normalized
mRNAseq_raw_counts
mRNAseq_RPKM
mRNAseq_RPKM_log2
mRNAseq_RPKM_Z_Score
uncv2.mRNAseq_raw_counts
uncv2.mRNAseq_RSEM_all
uncv2.mRNAseq_RSEM_normalized_log2
uncv2.mRNAseq_RSEM_Z_Score
Mutations
byGene
byLocation
Reverse Phase Protein Array
protein_exp_normalization
rppa
Filters
Search
ESR1|2099 (775)
ESYT3|83850 (772)
ETS2|2114 (775)
EVC|2121 (775)
EXD1|161829 (775)
ESR2|2100 (775)
ETAA1|54465 (775)
ETV1|2115 (775)
EVI2A|2123 (775)
EXD2|55218 (775)
ESRP1|54845 (775)
ETF1|2107 (775)
ETV2|2116 (774)
EVI2B|2124 (775)
EXD3|54932 (775)
ESRP2|80004 (775)
ETFA|2108 (775)
ETV3L|440695 (759)
EVI5L|115704 (775)
EXO1|9156 (775)
ESRRA|2101 (775)
ETFB|2109 (775)
ETV3|2117 (775)
EVI5|7813 (775)
EXOC1|55763 (775)
ESRRB|2103 (752)
ETFDH|2110 (775)
ETV4|2118 (775)
EVL|51466 (775)
EXOC2|55770 (775)
ESRRG|2104 (774)
ETHE1|23474 (775)
ETV5|2119 (775)
EVPLL|645027 (719)
EXOC3L2|90332 (775)
ESX1|80712 (760)
ETNK1|55500 (775)
ETV6|2120 (775)
EVPL|2125 (775)
EXOC3L|283849 (775)
ESYT1|23344 (775)
ETNK2|55224 (775)
ETV7|51513 (775)
EVX1|2128 (484)
EXOC3|11336 (775)
ESYT2|57488 (775)
ETS1|2113 (775)
EVC2|132884 (775)
EWSR1|2130 (775)
EXOC4|60412 (775)
BACK TO START
BACK 50
FORWARD 50
FORWARD 500
FORWARD 5000
Show All
Values > 30%
Values > 60%
Exceptions
SWITCH TO GENETIC SEARCH


Atlas Correlation Explorer
Atlas Correlation Explorer - Genetic Search View
Cancer Type : BRCA
1 Source Measures, 19196 Target Measures
1097 Samples after Filtering
Pause Evolution
Organisms
21132
66% Cached
1 to 10
Simple Leaders
+10
Selected Leader (775)
Coverage
ESR1|2099 (l) >>> C6orf97|80129
ESR1|2099 >>> CA12|771
ESR1|2099 >>> CT62|196993
ESR1|2099 (s) >>> XBP1|7494
ESR1|2099 >>> FLJ45983|399717
ESR1|2099 >>> FOXA1|3169
ESR1|2099 >>> ANXA9|8416 (s)
ESR1|2099 >>> C6orf211|79624
ESR1|2099 (s) >>> FSIP1|161835 (s)
ESR1|2099 >>> AFF3|3899
XBP1|7494
mRNAseq_RPKM_log2
Score : 0.6113023
3.52
2.250
-4.774
9.16
mRNAseq_RPKM_log2
ESR1|2099 (s)
SWITCH TO DATA SELECTION