Analysis of multiple protein marker panels: Optimizing sensitivity and selectivity

M. Athanas[1], A, Prakash[2], T. Rezai[2], B. Krastins[2], D. Sarracino[2], Kypros Nicolaides[3], Ramesh Kuppusamy[3], Mary F. Lopez[2]

[1]VAST Scientific, Cambridge, MA, USA, [2]Thermo Fisher Scientific, BRIMS, Cambridge, MA, USA, [3]Fetal Medicine Foundation, London, UK

Overview

Purpose: To demonstrate optimized sensitivity and selectivity from combining multiple marker candidates to build a panel of candidate biomarkers.

Methods: A study comprised of maternal serum samples extracted from 40 mothers with normal pregnancy and 40 mother giving birth to Trisomy 21 afflicted children. Data were process on a Thermo LTQ OrbiTrap XL and analyzed with SIEVE.

Results: ROC analysis in discovery is a robust discriminate in cohort studies. Combining multiple markers in some circumstances may provide a more revealing discriminate.

Introduction

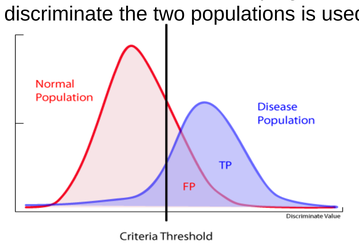
Proteomic discovery experiments are rapidly generating lists of putative biomarkers for diseases and pathologies. Verification of these markers in multiplexed assays poses a statistical challenge as traditional ROC (Receiver Operation Characteristic) curves used to calculate the sensitivity and specificity of a diagnostic or predictive assay are based on single markers. The ability to combine quantitative information from several markers could potentially improve the diagnostic accuracy of existing tests and facilitate the development of new tests. However, standardized approaches to representing panels of markers remains controversial.

In this analysis, we apply proteomics and mass spectrometry techniques for the discovery of new putative biomarkers for Trisomy 21 in first trimester maternal serum. Maternal serum samples from Trisomy 21 and normal first trimester pregnancies were provided by the Fetal Medicine Foundation and collected from study participants with full consent and approval. High resolution LC-MS/MS analysis was carried out on an LTQ-Orbitrap XL mass spectrometer (ThermoFisher Scientific).

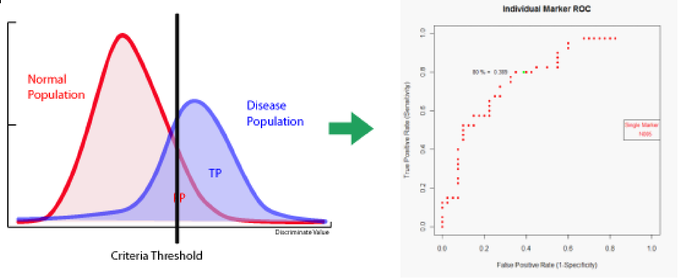
Marker Panels and ROCs

Single Markers and ROC Curve Plots

When considering the outcome of a specific measurement between disease and normal populations, typically there is some degree of overlap and the two populations are not perfectly separated. In a traditional ROC analysis, calculating the *true positive* (**TP**) rate (sensitivity) and the *false positive (***FP**) rate (100-specificity) is one way of assessing the efficacy of the specific measurement to differentiate the two populations. In this case, accounting TPs and FPs for the measurement while sweeping the cut-off point or threshold used to discriminate the two populations is used to construct a ROC curve.



For the example illustrated above, a traditional ROC plot is constructed by tabulating the FPs and TPs as the criteria threshold is swept across both of the cur



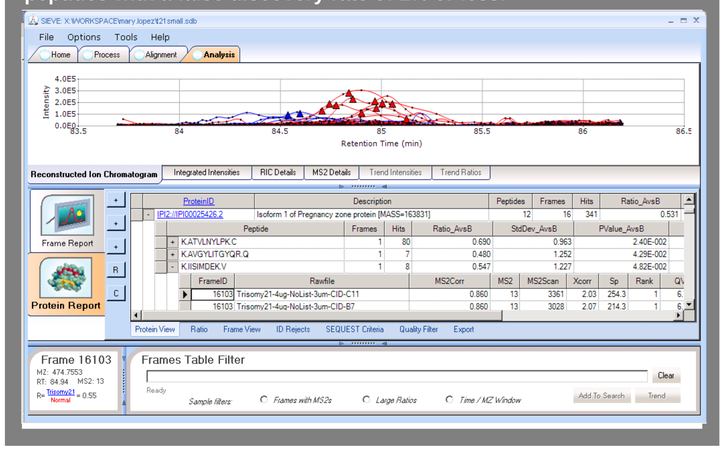
In a biological system, there may be multiple marker candidates that work in tandem with their own individual discriminating capability. The overall discriminating capability may be improved if a panel of markers were used instead of a single marker.

Multiple Markers and ROC Scatter Plots

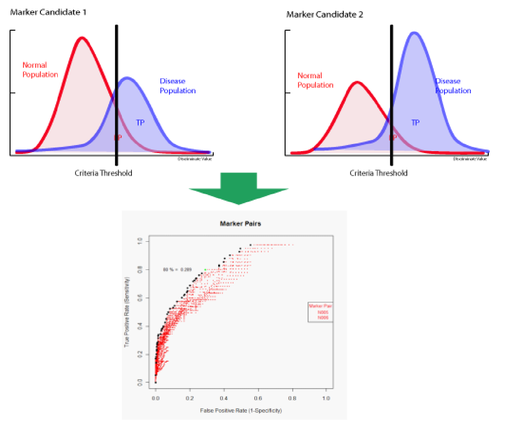
Two or more marker candidates can be combined using the Marker Multiplex ROC described below. In this case, the TP and FP is derived by the combined probability for each criteria threshold for each marker. That is, for a given set of criteria threshold {t}, the effective true positive value is derived from the product of each marker's true positive value evaluated at a select criteria threshold. Similarly, the same applies to the false positive value:

$$TP_{\{t\}} = \prod_{All\ Markers}^{t} TP_{\{i\}} \qquad FP_{\{t\}} = \prod_{All\ Markers}^{t} FP_{\{i\}}$$

FIGURE 1. Chromatographic alignment, feature detection, and feature identification was performed by SIEVE, a label-free LC-GC/MS statistical analysis platform from Thermo Fisher Scientific. The Percolator (2) algorithm incorporated into SIEVE selects only peptides with a false discovery rate of 2% or less.



For the example illustrated above, a traditional ROC plot is constructed by tabulating the FPs and TPs as the criteria threshold is swept across both of the curves.



The most effective discriminating power of the combined marker set is found as the top-leftmost edge of the scatter plot shown above as black points. An overall efficacy of the marker panel can be expressed as the area under the curve obtained by joining the points along the top-leftmost edge of the distribution.

Ascertaining and assessing candidate markers in discovery experiments in which measurements are obtained across multiple classes (biological replicates, dosage studies, time series, trend analyses, etc.) provides a more powerful differentiator instead of fold change or ratio comparison.

Data Analysis Procedure

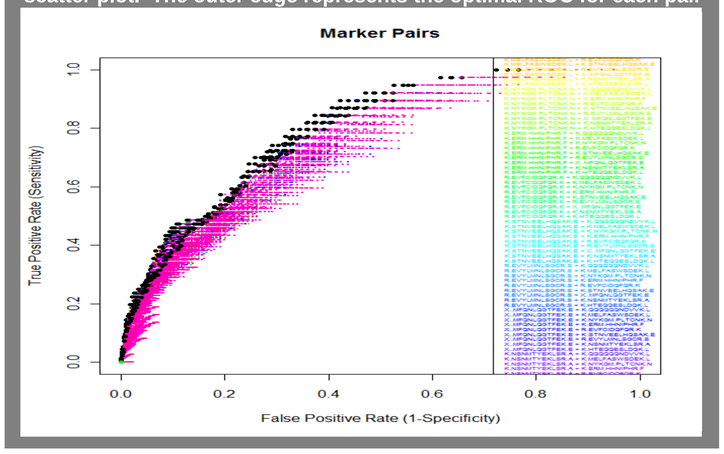
A SIEVE "Control vs Treatment" experiment was constructed from the 40 normal and 40 trimsomy21 samples. 20k features were produced after chromatographic alignment and framing. Identification of frames is handled by SEQUEST and peptide rescoring is processed by Percolator [2]. 5671 distinct peptides were found with a false discovery rate set to 2% or less. These peptides non-uniquely were assigned to 4221 proteins.

ROC analysis was performed on every peptide. The ROC area-under-the-curve (AUC) is used to rank the most discriminating peptides.

FIGURE 2. Top 25 peptides ranked by ROC area-under-the-curve. Also shown is the Percolator False Discovery Rate (FDR).

#	Peptide	AUC	FDR	Description
1	K.QQQQQQNDVVK.L	0.718	1.68E-11	NP_006852.1| RAB35, member RAS oncogene family [Homo sapiens] [MASS=23025]
2	K.MELFASWSDEK.L	0.708	1.66E-10	NP_543024.1| hypothetical protein LOC140894 [Homo sapiens] [MASS=67066]
3	K.NYKGM*PLTCNK.N	0.708	1.92E-11	NP_872415.1| zinc finger protein 677 [Homo sapiens] [MASS=67996]
4	K.ERM*HHNIPHR.F	0.708	4.13E-11	NP_009105.1| citron [Homo sapiens] [MASS=231430]
5	R.EVFCIDQFQR.K	0.708	8.67E-12	NP_079279.3| a disintegrin-like and metalloprotease with thrombospondin type 1 motifs 20 [Homo sapiens] [MASS=214719]
6	K.STNVEELHQSAK.E	0.707	6.98E-11	NP_787070.1| hypothetical protein LOC144608 [Homo sapiens] [MASS=27654]
7	R.EVYLMNLSGCR.S	0.706	2.09E-10	NP_031363.2| deleted in lung and esophageal cancer 1 isoform DLEC1-S3 [Homo sapiens] [MASS=197896]
8	-.MFQNLQGTFEK.E	0.706	2.09E-10	NP_997318.1| hypothetical protein LOC400073 [Homo sapiens] [MASS=14959]
9	K.NSNMTYEKLSR.A	0.703	5.38E-13	NP_004424.2| E74-like factor 3 (ets domain transcription factor, epithelial-specific) [Homo sapiens] [MASS=41455]
10	K.HTEQQESLDQK.L	0.703	8.02E-13	NP_443723.2| ankyrin repeat domain 30A [Homo sapiens] [MASS=152673]
11	K.HXVLSAPCTDTK.G	0.703	4.48E-12	XP_939782.2| PREDICTED: similar to Zinc finger protein 443 (Kruppel-type zinc finger protein ZK1) isoform 2 [Homo sapiens] [MASS=130674]
12	R.GAPDLPRGGCEGR.R	0.703	4.13E-11	XP_937902.2| PREDICTED: hypothetical protein [Homo sapiens] [MASS=82387]
13	K.PAALPEKEGGAAK.S	0.703	1.87E-12	XP_001133680.1| PREDICTED: hypothetical protein [Homo sapiens] [MASS=15595]
14	R.KELLGM*CFGDR.Y	0.702	1.73E-12	NP_722561.1| G protein-coupled receptor 161 isoform 2 [Homo sapiens] [MASS=58559]
15	K.LDAHVERVDCK.A	0.702	2.09E-10	NP_000457.1| peroxin1 [Homo sapiens] [MASS=142866]
16	K.EFNFNLGEMNK.S	0.698	2.09E-10	NP_653200.2| non-imprinted in Prader-Willi/Angelman syndrome 1 [Homo sapiens] [MASS=34562]
17	R.MPNLM*GHENQR.E	0.696	8.63E-12	NP_009128.1| frizzled 10 [Homo sapiens] [MASS=65335]
18	R.GFAEHGKQQQGR.E	0.696	6.98E-11	NP_056114.1| hypothetical protein LOC23351 [Homo sapiens] [MASS=74533]
19	K.IDRM*MEAFASR.Y	0.696	5.42E-11	NP_004218.1| pleckstrin homology, Sec7 and coiled/coil domains 3 [Homo sapiens] [MASS=46291]
20	K.IGSSLPQDDDAPK.K	0.696	1.73E-12	NP_996744.1| transforming, acidic coiled-coil containing protein 2 isoform a [Homo sapiens] [MASS=309395]
21	K.RQNEGNPERDK.E	0.696	1.21E-10	NP_958848.1| periphilin 1 isoform 4 [Homo sapiens] [MASS=36952]
22	K.LM*DHVGTEPSIK.E	0.696	8.61E-12	NP_002942.2| ribophorin II precursor [Homo sapiens] [MASS=69283]
23	K.YFSAAEYHAQR.L	0.696	1.66E-10	NP_849149.2| IQ motif and ubiquitin domain containing [Homo sapiens] [MASS=92537]
24	R.HM*VVHSGDGPYK.C	0.696	1.92E-11	NP_689475.1| zinc finger protein 439 [Homo sapiens] [MASS=58395]
25	K.SELLPAGWNNNK.D	0.696	1.21E-10	NP_848693.2| proteasome inhibitor subunit 1 [Homo sapiens] [MASS=29817]

FIGURE 3. A ROC scatter plot of all possible pairs of the top 10 candidate peptide markers. The color points represent a possible FP and TP threshold value for a pair of peptides (legend on the right). The black points represent the outer edge of each individual ROC scatter plot. The outer edge represents the optimal ROC for each pair



The most effective discriminating power of the candidate marker set is found when constructing the multi-marker ROC. The multi-marker ROC is calculated for the top ten candidate markers. The Area-Under-the-Curve (AUC) of the outer edge of the ROC scatter plot was calculated for every peptide-peptide combination. The top 25 pairs are shown in Figure 4.

In the table Figure 4, the peptide pairs K.QQQQQQNDVVK.L (RAB35, member RAS oncogene family) and K.MELFASWSDEK.L (hypothetical protein LOC140894) and K.QQQQQQNDVVK.L and K.NYKGM*PLTCNK.N (zinc finger protein 677) scored significantly higher (12%) than individual peptide ROC AUC.

FIGURE 4. A multi-ROC scatter plot is processed for every peptide candidate pair. The Area-Under-the-Curve (AUC) is calculated from the outer most edge of the ROC scatter plot. The marker pair discriminating capability is captured by the AUC where larger values correspond to better selectivity and sensitivity.

Column1	AUC	Peptide Marker Pair
1	0.805	K.QQQQQQNDVVK.L + K.MELFASWSDEK.L
2	0.805	K.QQQQQQNDVVK.L + K.NYKGM.PLTCNK.N
3	0.803	K.QQQQQQNDVVK.L + K.ERM.HHNIPHR.F
4	0.805	K.QQQQQQNDVVK.L + R.EVFCIDQFQR.K
5	0.800	K.QQQQQQNDVVK.L + K.STNVEELHQSAK.E
6	0.800	K.QQQQQQNDVVK.L + R.EVYLMNLSGCR.S
7	0.801	K.QQQQQQNDVVK.L + X..MFQNLQGTFEK.E
8	0.797	K.QQQQQQNDVVK.L + K.NSNMTYEKLSR.A
9	0.798	K.QQQQQQNDVVK.L + K.HTEQQESLDQK.L
10	0.805	K.MELFASWSDEK.L + K.QQQQQQNDVVK.L
11	0.803	K.MELFASWSDEK.L + K.NYKGM.PLTCNK.N
12	0.801	K.MELFASWSDEK.L + K.ERM.HHNIPHR.F
13	0.803	K.MELFASWSDEK.L + R.EVFCIDQFQR.K
14	0.799	K.MELFASWSDEK.L + K.STNVEELHQSAK.E
15	0.799	K.MELFASWSDEK.L + R.EVYLMNLSGCR.S
16	0.799	K.MELFASWSDEK.L + X..MFQNLQGTFEK.E
17	0.797	K.MELFASWSDEK.L + K.NSNMTYEKLSR.A
18	0.797	K.MELFASWSDEK.L + K.HTEQQESLDQK.L
19	0.805	K.NYKGM.PLTCNK.N + K.QQQQQQNDVVK.L
20	0.803	K.NYKGM.PLTCNK.N + K.MELFASWSDEK.L
21	0.798	K.NYKGM.PLTCNK.N + K.ERM.HHNIPHR.F
22	0.797	K.NYKGM.PLTCNK.N + R.EVFCIDQFQR.K
23	0.796	K.NYKGM.PLTCNK.N + K.STNVEELHQSAK.E
24	0.796	K.NYKGM.PLTCNK.N + R.EVYLMNLSGCR.S
25	0.796	K.NYKGM.PLTCNK.N + X..MFQNLQGTFEK.E

Conclusion

We have demonstrated a workflow for a two class study (Normal, Trisomy21) for the discovery of biomarker candidates using the ROC area under the curve as a candidate discriminate. We conclude:

- ROC analysis at the discovery phase of an experiment is a powerful way of identifying robust candidate biomarkers.
- Multiple marker distinguishing capability can be significantly enhance when compared to single marker capability.
- The significance of triplets and higher order marker combinations is diminished.

References

1. M, Lopez, R. Kuppusamy, D. Sarracino, A. Prakash, M. Athanas, B. Krastins, T. Rezai, J. Sutton, S. Peterman, and K. Nicolaides; *Discovery and targeted SRM assay development of first-trimester peptide biomarker candidates for Trisomy 21 in maternal blood*, submitted for publication
2. Lukas Käll, Jesse Canterbury, Jason Weston, William Stafford Noble and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets Nature Methods 4:923 – 925, November 2007