DART-ID Increases Proteome Coverage with Data-Driven Alignment of Retention Times

Albert Chen[1], Alexander Franks[2], Nikolai Slavov[1]

1 Departments of Biology, Bioengineering, Northeastern University, Boston, MA 02115, USA
2 Department of Statistics and Applied Probability, UC Santa Barbara, CA 93106, USA

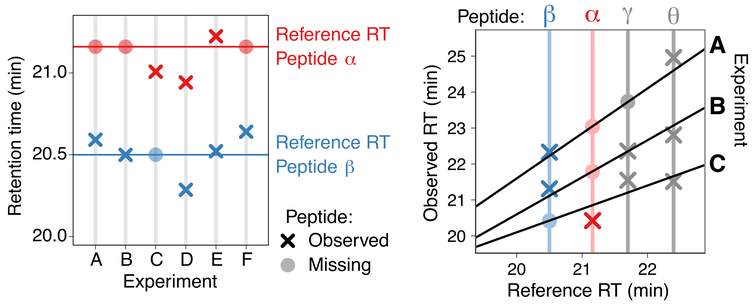
Slavov Laboratory
Quantitative Biology

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
DOI: 10.1101/399121

Summary

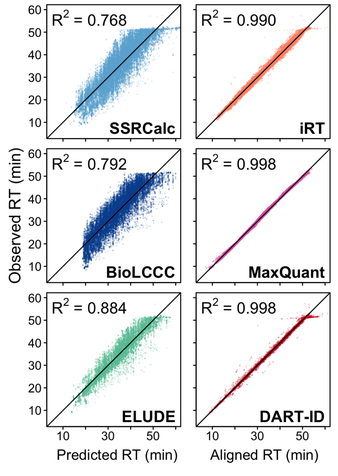
- Retention time (RT) is an informative feature that can be remarkably consistent across LC runs with the same sample and same experimental conditions.
- Data-driven alignments of peptide RTs across experiments create robust inferences of peptide RTs with RT distributions.
- Applying inferred and observed RTs within a principled Bayesian framework greatly increases the coverage of single cell proteomes.

Global RT Alignment Provides Robust RT Inferences Across Experiments



- Latent Reference RT (μ) inferred for each peptide sequence.
- Shifts between experiments modeled as monotonic transformation of μ.
- Optimization robust to large amounts of missing data

Global RT Alignment Results in Residuals < 1 min for 60 min LC Runs

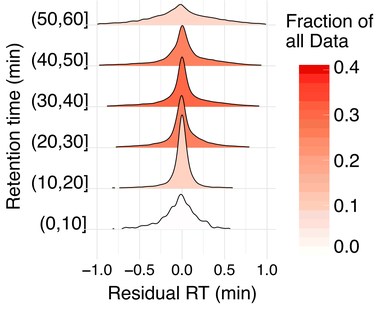


Summary statistics for residual RT (ΔRT)

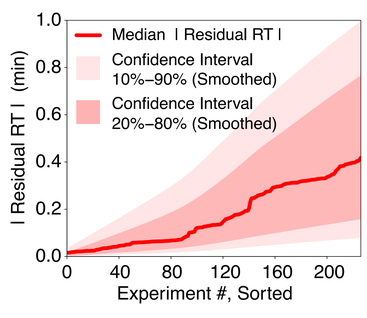
	Method	Mean|ΔRT|	Median|ΔRT|
Prediction	SSRCalc	3.77	3.19
	BioLCCC	3.44	2.91
	ELUDE	2.51	2.02
Alignment	iRT	0.693	0.537
	MaxQuant	0.329	0.243
	DART-ID	0.155	0.044

- 46 LC-MS/MS runs, 60 min each
- Residual RT = Observed RT - Predicted RT
- DART-ID shows smallest average deviation of 0.044 min (2.6 seconds)

Residual RT increases with time



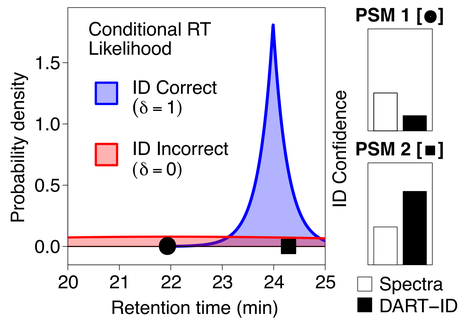
Residual RT varies by experiment



Bayesian Framework for Updating ID Confidence

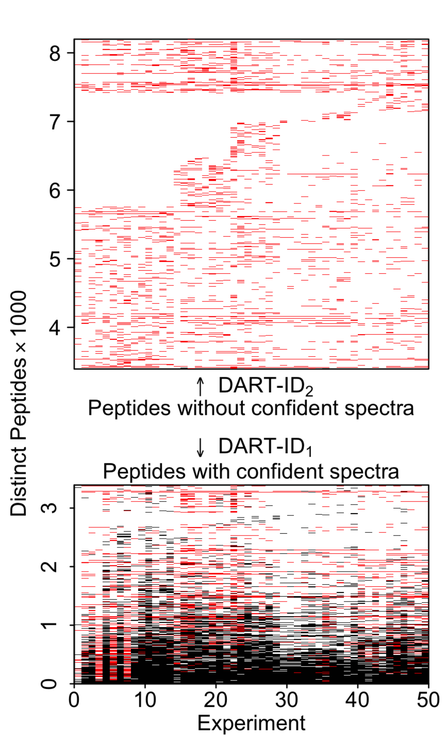
$$P(\text{ ID correct} \mid \text{RT}) = \frac{P(\text{ RT} \mid \text{ID correct}) \times P(\text{ ID correct})}{P(\text{ RT})}$$

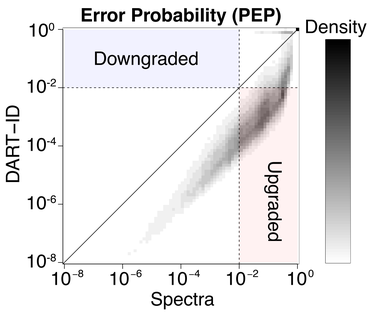
- $P(\text{ ID correct} \mid \text{RT})$ - Posterior probability that the PSM is assigned to the right sequence, given its RT. AKA "DART-ID" confidence.
- $P(\text{ RT} \mid \text{ID correct})$ - Conditional likelihood of the PSM's RT if its assigned sequence is correct. Estimated by evaluating inferred RT distribution of the sequence at the observed RT.
- $P(\text{ RT})$ - Marginal likelihood for observing the RT. Estimated as the sum of the probabilities that the PSM's assigned sequence is correct and incorrect.

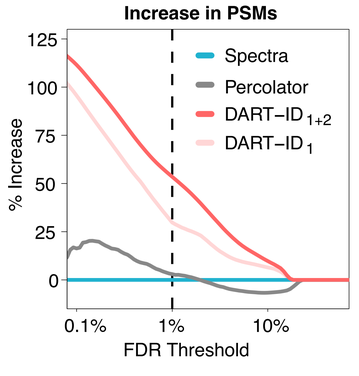


- Mean of inferred RT distribution is the transformed reference RT
- Inferred RT distribution variance experiment-specific
- Null distribution - empirical distribution of all RTs
- Case 1 - confidence upgraded
- Case 2 - confidence downgraded

ID Confidence Update Yields 50% More Peptides



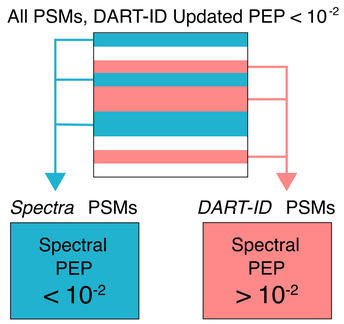




- At FDR threshold of 1%, up to 50% more PSMs identified
- Same total # of false discoveries after confidence update
- Also boost PSMs without confident spectra - need further validation

Not Quantified
Quantified (Spectra PEP < 1%)
Quantified (DART-ID PEP < 1%)

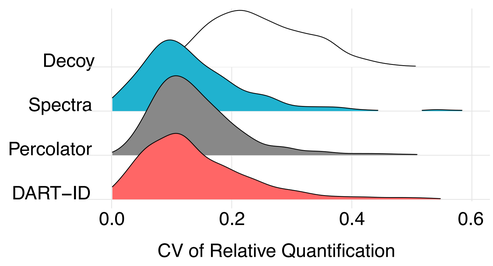
Validating New Identifications

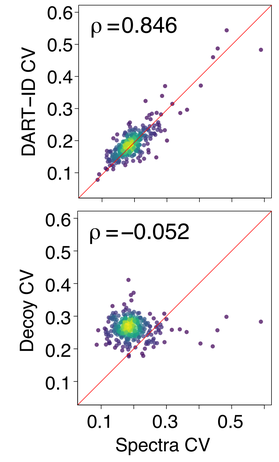


Split data into disjoint sets

- **Spectra** - PSMs filtered with PEP < 1%.
- **DART-ID** - PSMs filtered with PEP > 1% and updated PEP < 1%. This set of new observations is *disjoint* from Spectra PSM.
- **Percolator** - PSMs filtered with PEP > 1% and updated PEP < 1%. Also *disjoint* from Spectra PSMs.
- **Decoy** - PSMs filtered with PEP < 1%, and randomly sampled to create a decoy protein.

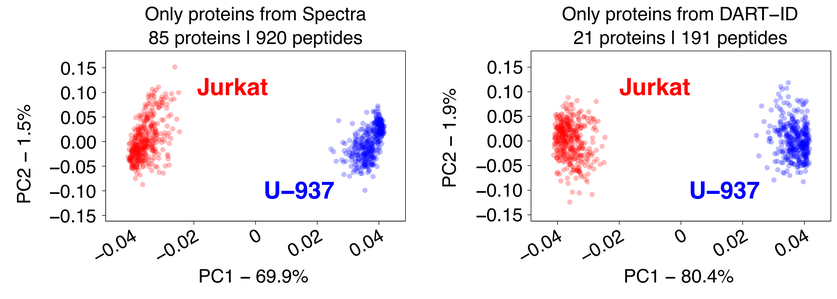
DART-ID PSMs Give Consistent Protein Quantification





- Coefficient of Variation (CV) = Standard deviation (σ) / mean (μ)
- Upgraded PSMs agree with previously confident PSMs, and are distinct from the decoy proteins.
- Degree of consistency preserved in proteins quantified from DART-ID PSMs alone

DART-ID Proteins Separate Cell Types



- Separation of the proteomes of 375 single cell samples each of 2 blood cancer cell lines, Jurkat and U-937 by Principal Component Analysis (PCA)

Conclusion

DART-ID takes advantage of reproducible retention times for peptide sequences within sets of LC runs to greatly increase the coverage of single cell proteomes.

Global alignment method provides more robust estimates of RT. The more consistent experiment RTs are, the more powerful the added RT evidence is.

Acknowledgements

We thank members of the Slavov laboratory for discussions and feedback. This work was funded by the National Institute of Health to N.S. under Project Number 1DP2GM123497-01.