DART-ID: Data-Driven Alignment of Retention Times for Peptide Identification Increases Peptide Coverage by > 85%

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

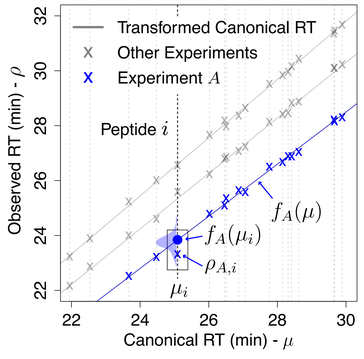
1 Departments of Biology, Bioengineering, Northeastern University, Boston, MA 02115, USA, **2** Department of Statistics and Applied Probability, UC Santa Barbara, CA 93106, USA

Slavov Laboratory
Quantitative Biology

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

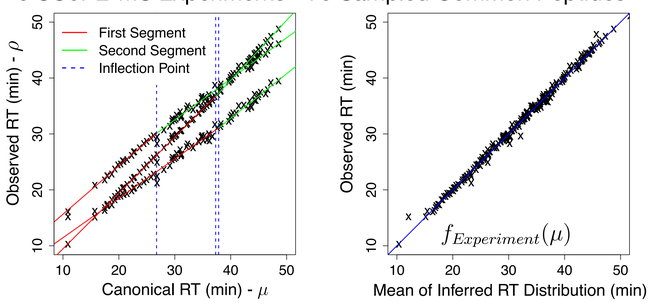


- Canonical RT (μ) for each peptide sequence.
- RT shifts between experiments modeled as monotonic transformation of canonical RTs.
- Singular error term for entire alignment.
- Optimization method adjusts μ and f_A to minimize single error term.

Two-Segment Linear Fit

- Two-segment fit captures more variation. More segments or non-linear models possible, as long as monotonicity holds.

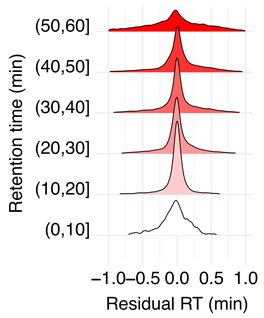
3 SCoPE-MS Experiments - 75 Sampled Common Peptides



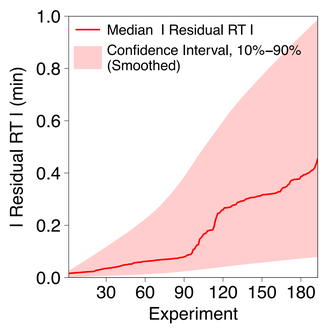
RT Alignment Residuals < 1 min. for 60 min LC Runs

Residual RT = Observed RT - Inferred RT

Residual RT increases with time



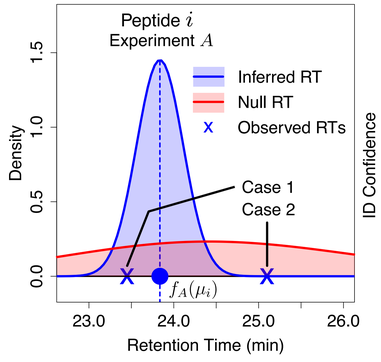
Residual RT varies by experiment

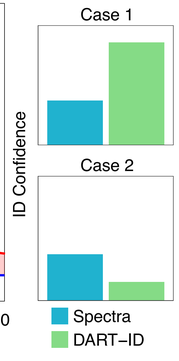


Bayesian Framework for Updating ID Confidence

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

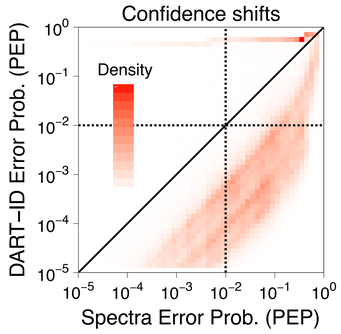
- $P(\text{ ID correct} \mid \text{RT})$ - Posterior probability that the PSM is assigned to the right sequence, given its RT. AKA "DART-ID" confidence.
- $P(\text{ RT} \mid \text{ID correct})$ - Conditional likelihood of the PSM's RT if its assigned sequence is correct. Estimated by evaluating inferred 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.

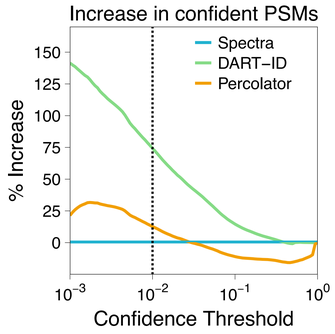




- Mean of inferred RT distribution is the transformed canonical RT
- Inferred RT distribution variance experiment-specific
- Null distribution - empirical distribution of all RTs
- Case 1 - confidence upgraded
- Case 2 - confidence downgraded

ID Confidence Update Yields 75% More PSMs

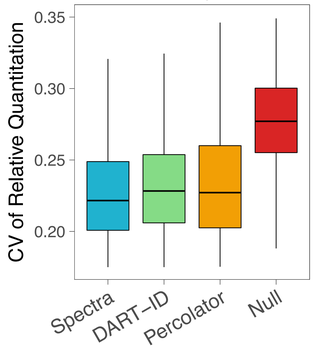




- 45 deg. line separates PSM confidence upgrades (below) and downgrades (above).
- At confidence threshold of 0.01 (1%), 75% more PSMs selected.

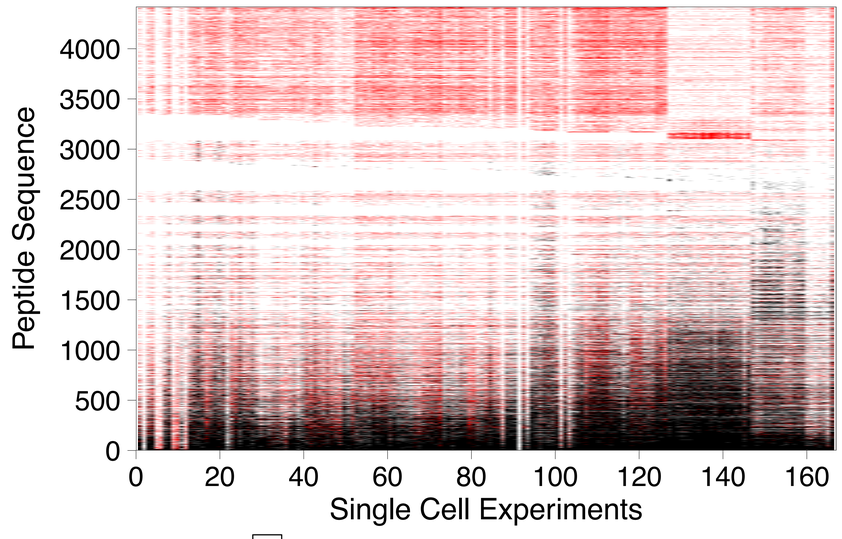
Newly Upgraded IDs Consistent with Existing IDs

Consistency of Protein Quantitation in SCoPE-MS Experiments

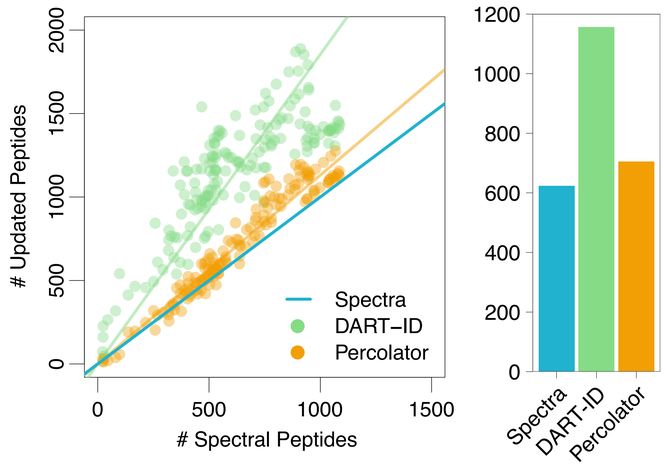


- Coefficient of variation (CV: σ / μ) of relative quantitation of PSMs within proteins (n=1590).
- **Spectra** - PSMs filtered with PEP < 1%.
- **DART-ID** - PSMs filtered with PEP > 1% and updated PEP < 1%. This set of new observations is *disjoint* from Spectra PSMs.
- **Percolator** - PSMs filtered 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.
- Upgraded PSMs agree with previously confident PSMs, and are distinct from the decoy proteins.

Substantial Increase in Proteome Coverage



Peptides Quantified Per Experiment (PEP < 1%)



- At confidence threshold of 1%, on average > 85% of quantified peptides per experiment.

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.