



Combining MS/MS Sequence and Spectral Library Search Algorithms

Lewis Geer, NCBI/NLM/NIH

Paul Rudnick and Stephen Stein, NIST

Overview of search methods

CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Sequence search

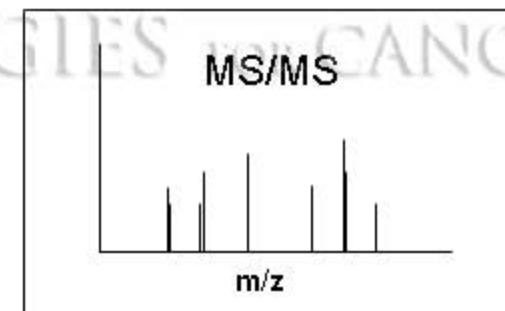
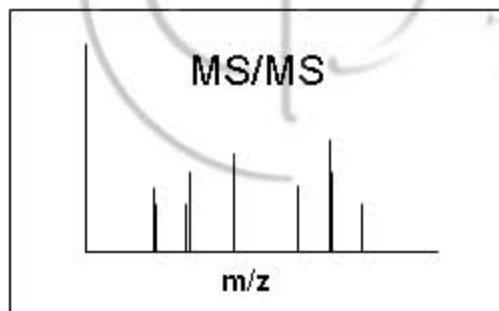
Experiment



Sequence Library

```
>gi|129369|sp|P04637|P53_HUMAN Cellular tumor antigen p53  
MEEPQSDPSVEPPLSQETFSDLWKLIPENNVLSPILPSQAMDDLMLSPDDIEQWFTE  
PPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYS  
CPVQLWVDSTPPPGTRVRAMAIIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRV
```

Theoretical digest



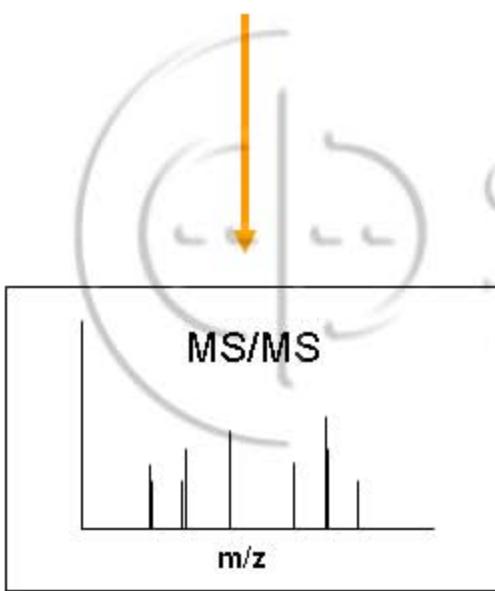
Match precursor ion (optionally)

Match product ions

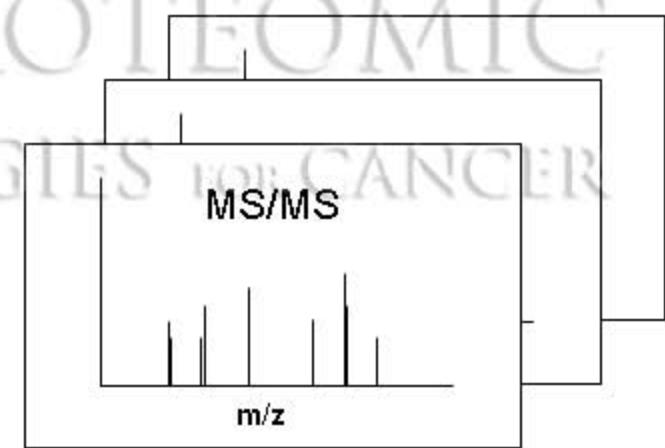
Score

Spectral library search

Experiment



Spectral Library

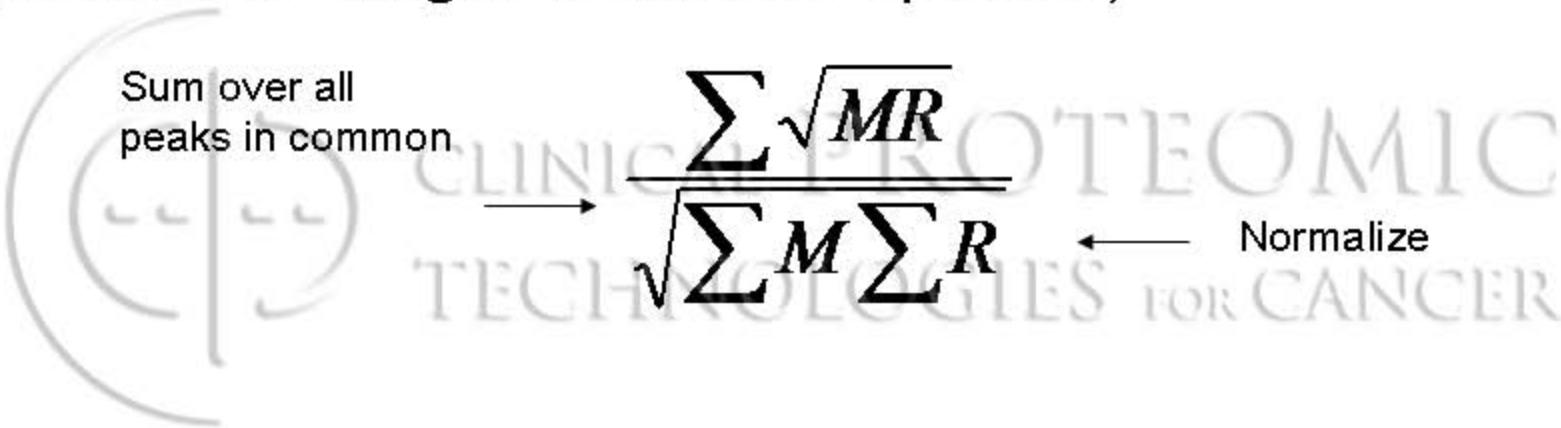


Match precursor ion (optionally)

Match product ions

Score

Spectral library search score based on ‘Dot Product’ (cosine of ‘angle’ between spectra)


$$\frac{\sum \sqrt{MR}}{\sqrt{\sum M \sum R}}$$

Sum over all peaks in common

Normalize

- $M = f(\text{Abundance})$ Peak in Measured Spectrum
- $R = f(\text{Abundance})$ Peak in Reference Spectrum
- $f(\text{Abundance})$
 - Weight as you like

Building a spectral library

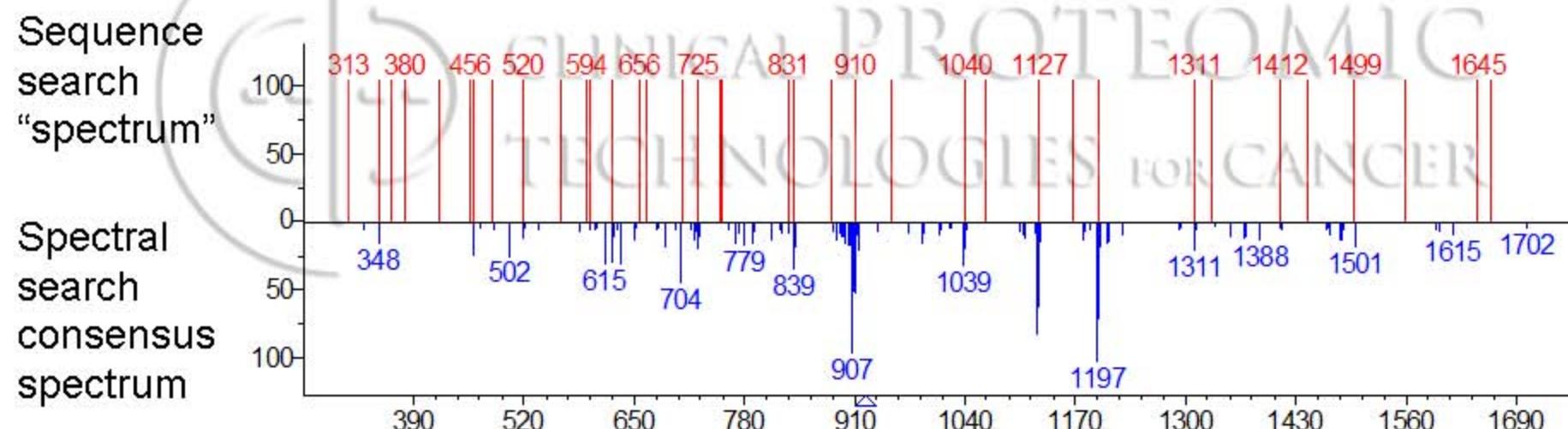
- Extract identified spectra from sequence search
 - Multiple search engines
 - Instrument-class specific
 - Use reversed sequence library
- Create 'consensus' spectra
 - Two or more matching spectra, also save best
- Assign probability of being correct
 - Refine confidence
- Create searchable spectral library
 - Resolve conflicts, add annotation



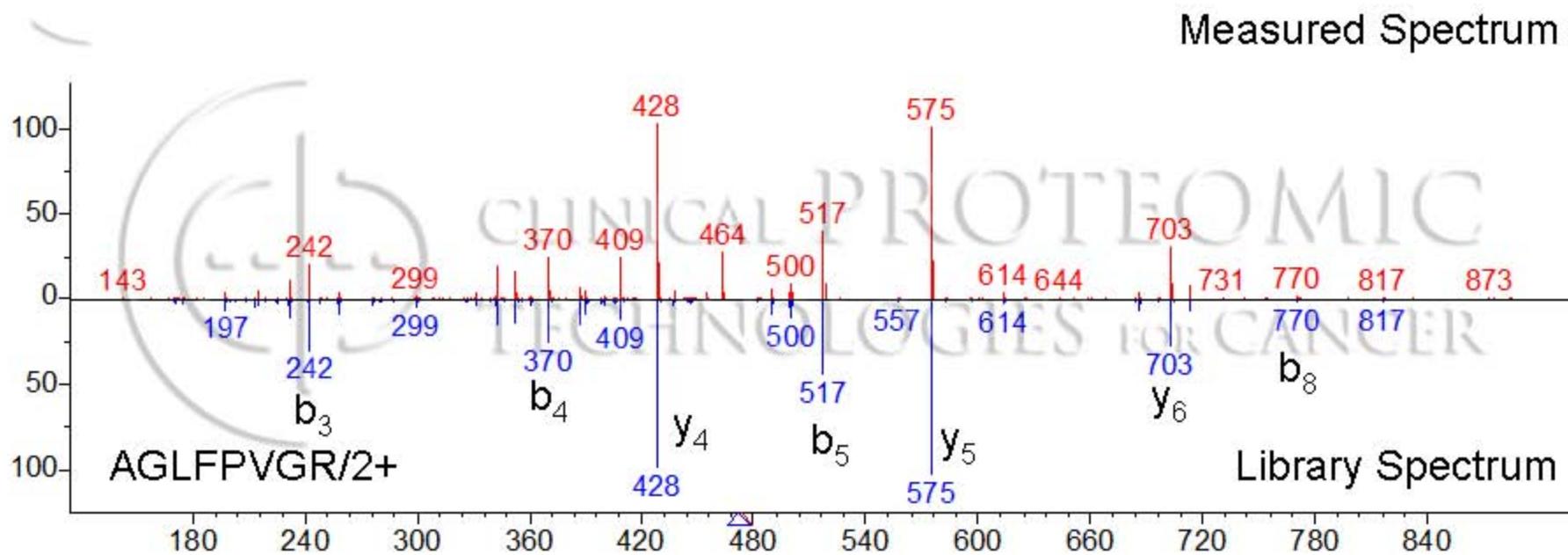
Why combine the search methods?

CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

“Reference” spectra are different depending on search method



Spectral library match



- scoring includes intensity of reference spectrum
- includes “non-canonical” peaks
- does not include canonical peaks not present

Matching strengths to weaknesses: coverage and sensitivity

Sequence searching

■ Strengths

- Search space is determined by sequence library
- Required to identify new spectra

■ Weaknesses

- limited use of intensity, non-canonical ion species, and ions that are not present
- large search space reduces sensitivity and speed

Spectral library searching

■ Strengths

- Use of peak intensities improves scoring sensitivity and specificity
- Uses comprehensive annotation of ions in library spectra
- limits search space to previously identified peptides

■ Weaknesses

- Search space limited to spectral library



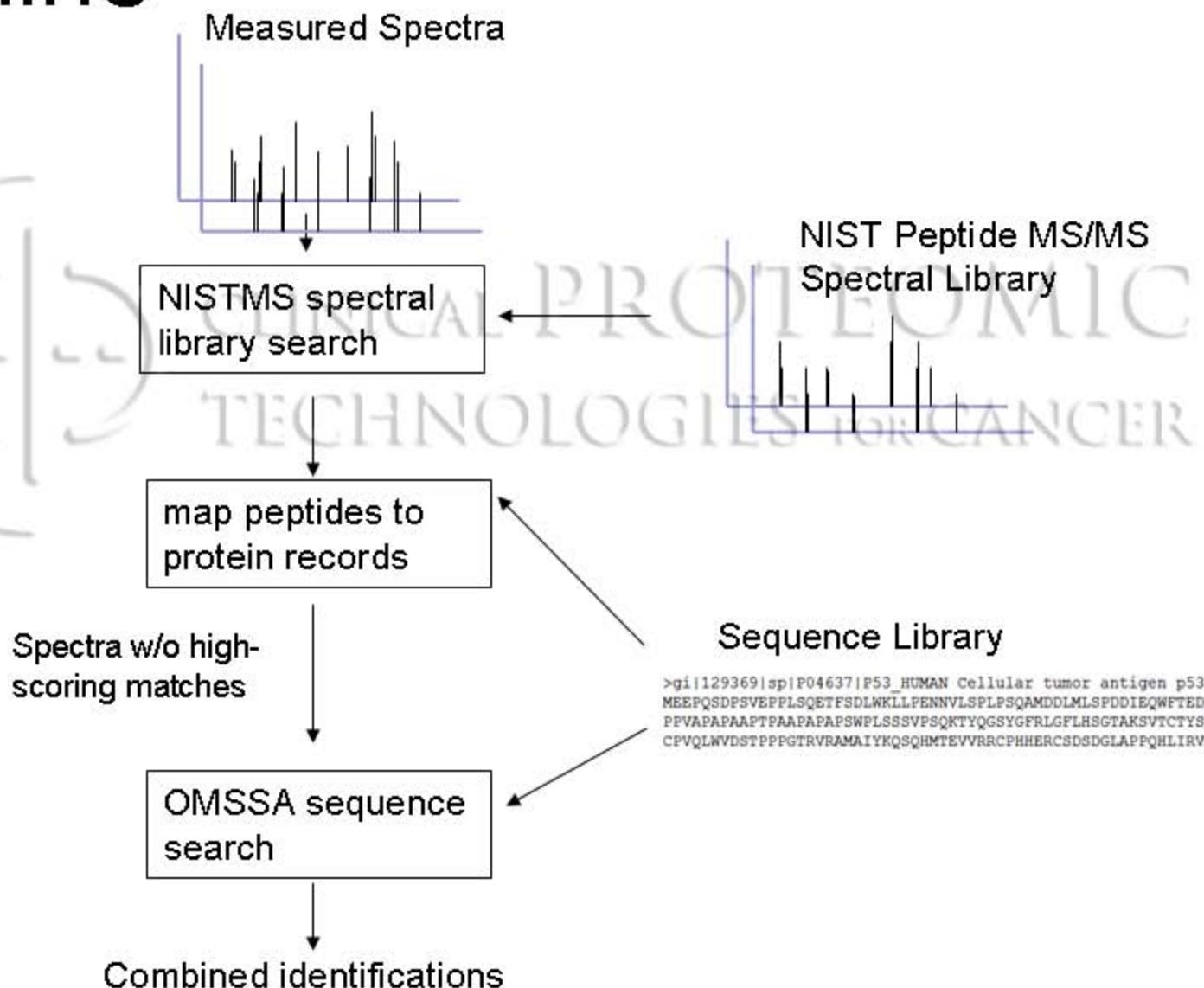
CLINICAL PROTEOMIC

How can the methods be combined?

Issues in combining the search methods

- Create a pipeline
- Combine the scores
 - plot the scores for each algorithm versus a measured FDR
 - match the FDRs
 - convert NIST score to OMSSA compatible e-value score

Pipeline



Combined score: FDR calculation

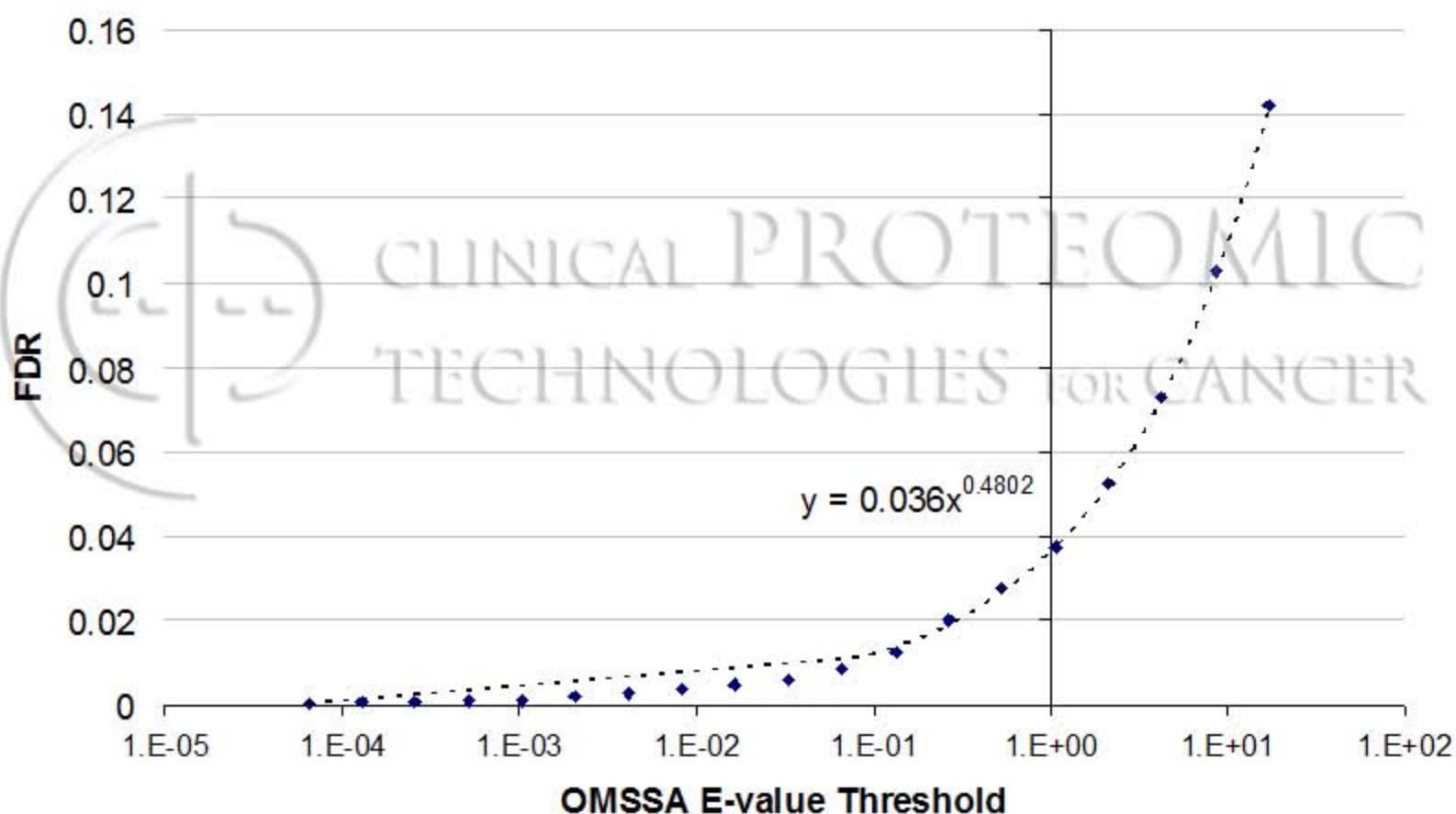
- Start with a large dataset - ~100,000 LTQ spectra from a 2-D separation, 'fresh frozen', human breast tissue isolated by LCM
- Use decoy databases to calculate FDR
- *human_decoy*: fly, yeast, *D. radiodurans*

library sizes	
seq. forward	54478
seq. decoy	33064
spec. forward	184383
spec. decoy	149633

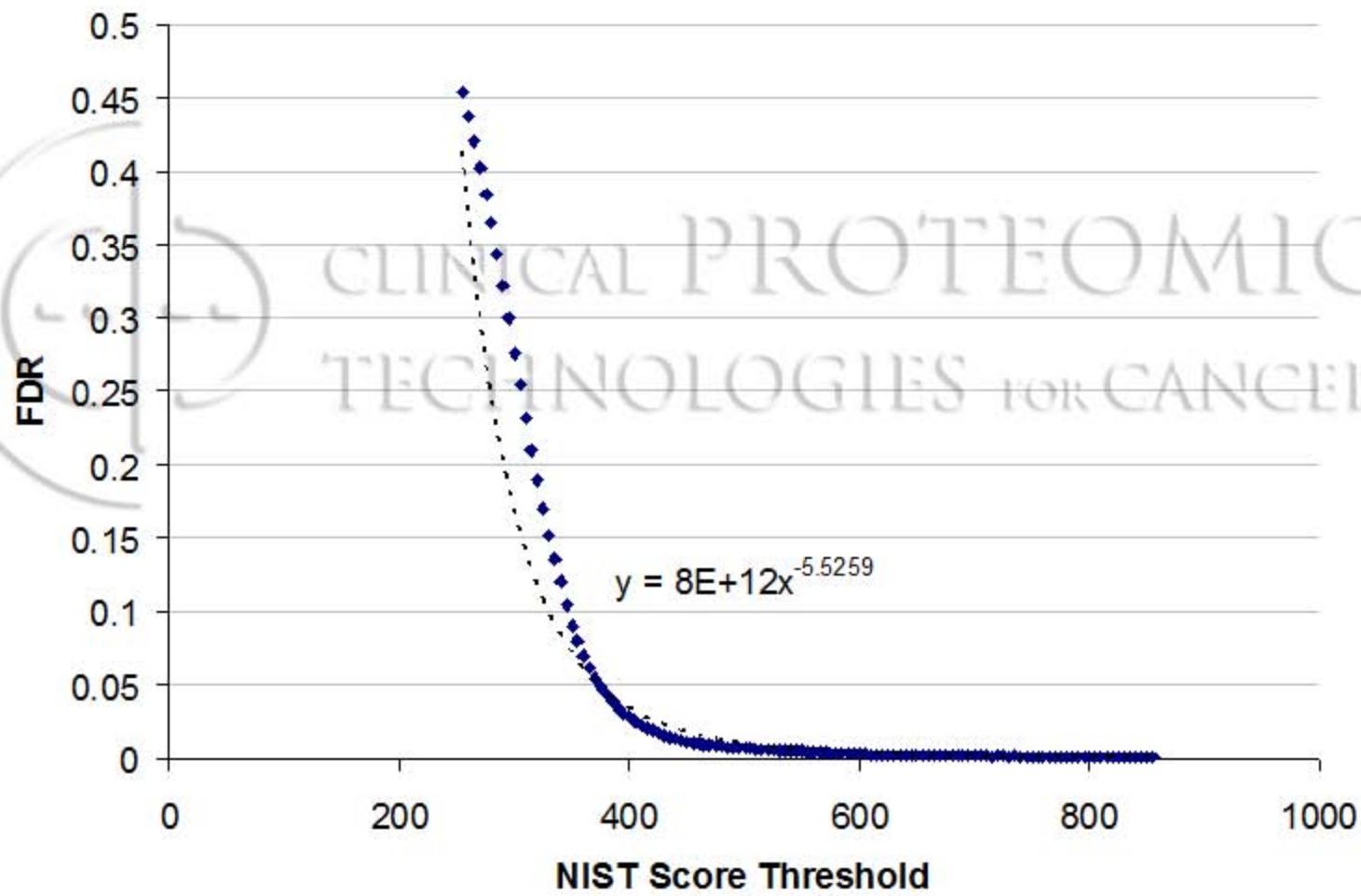
Combined score: FDR calculation

- Use paired forward and decoy searches
- Exclude from false matches:
 - Peptide isobars matching forward sequences [e.g. (Q|K), (I|L)]
 - Matches scoring higher against forward library, reduce scoring “boosts” from single match spectra

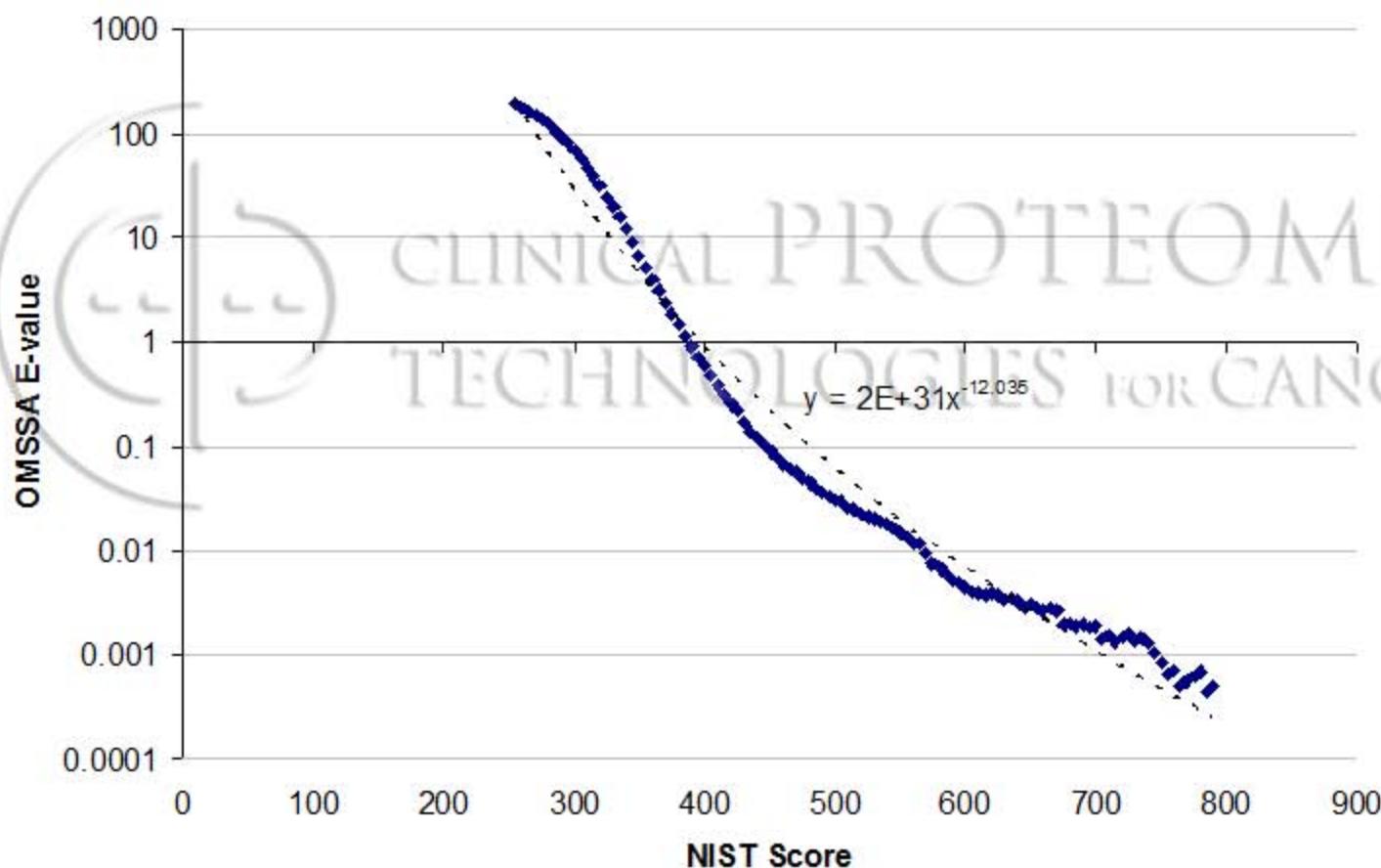
Transformation of OMSSA E-values to FDR



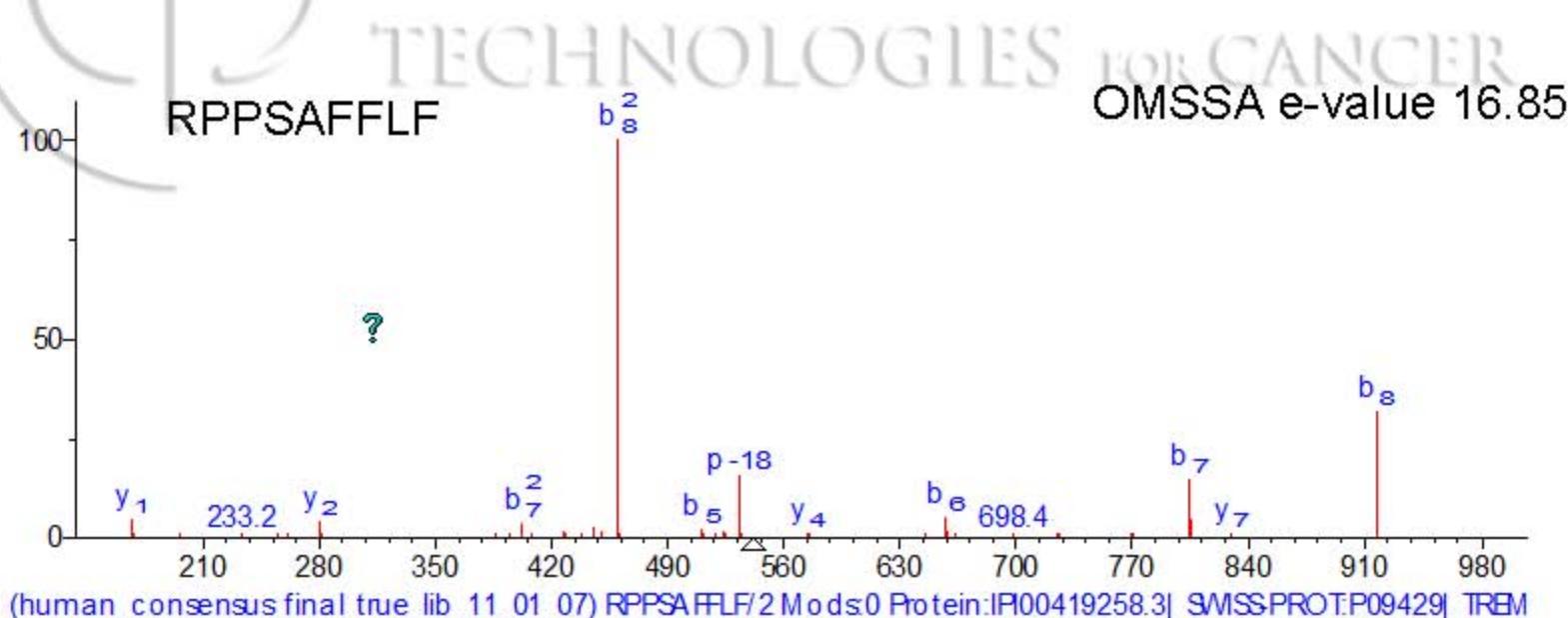
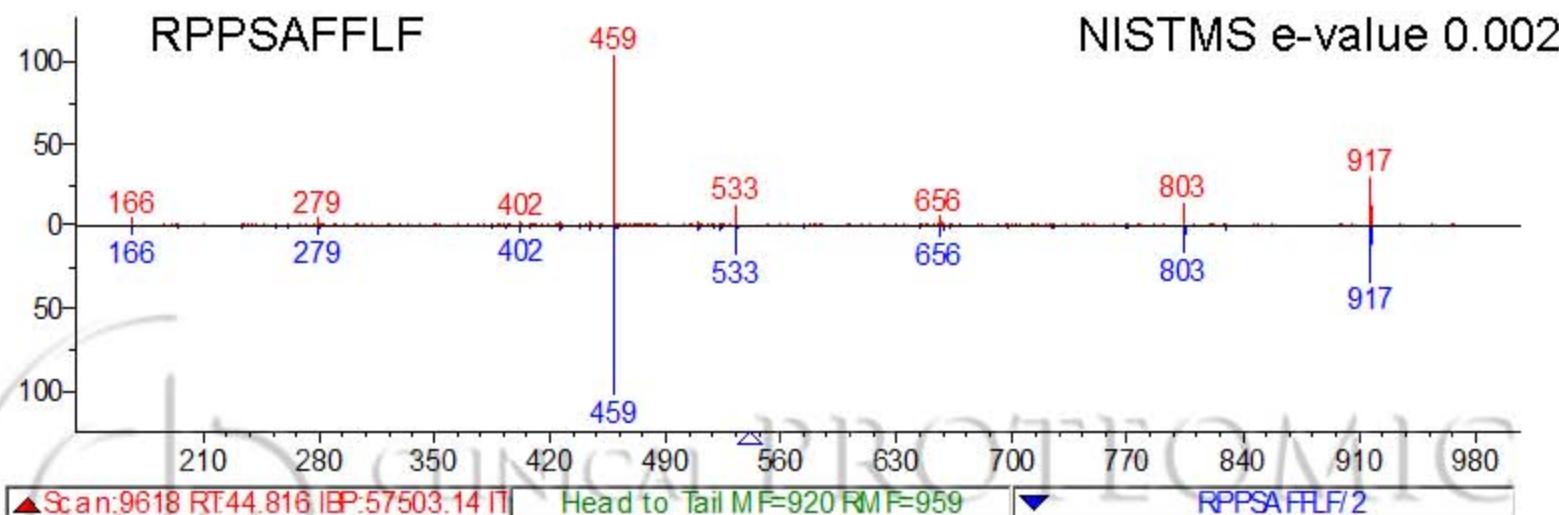
Transformation of NIST Score to FDR



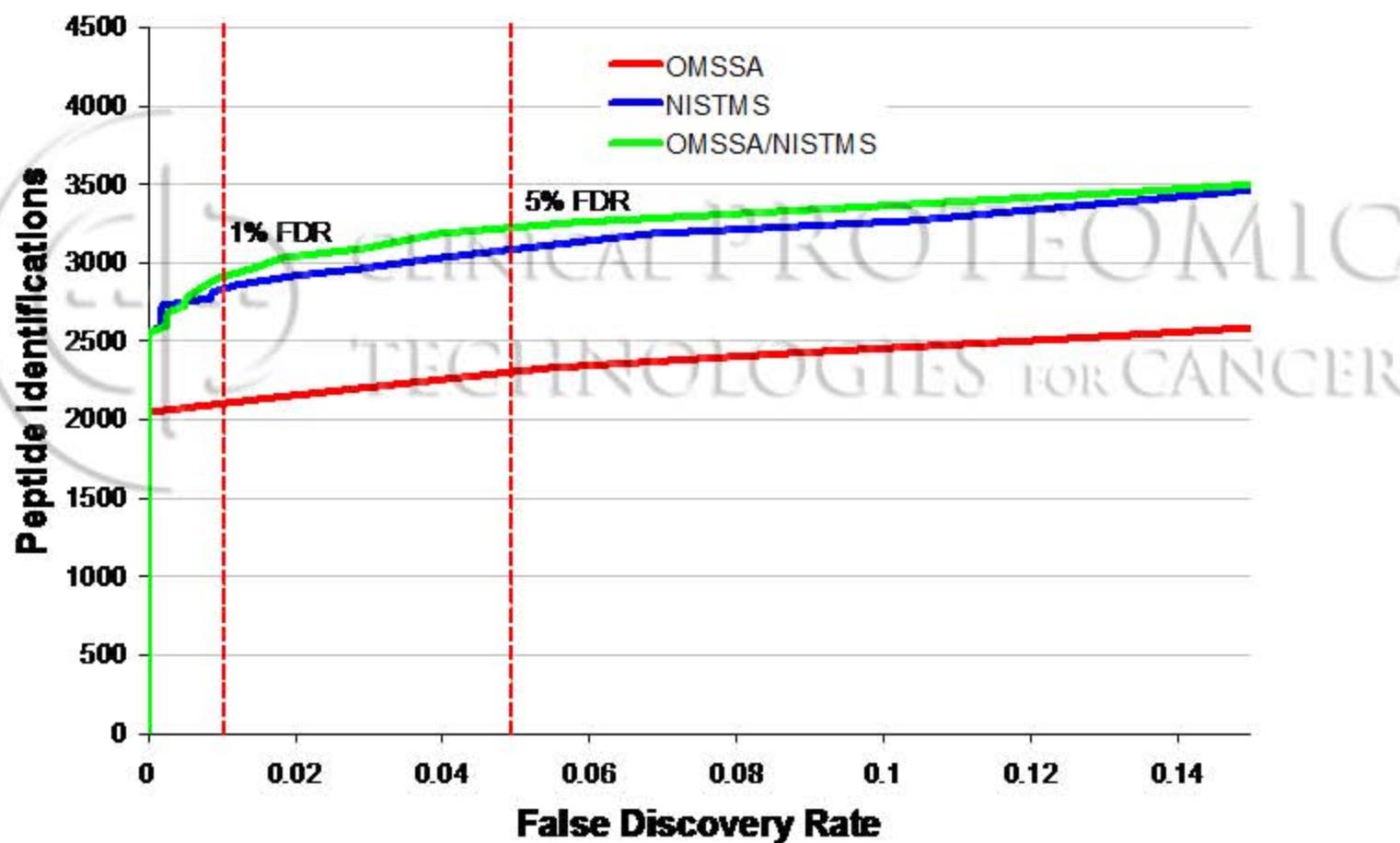
Conversion between NIST Score and OMSSA E-value equivalents



Transform scaled by relative size of library



Performance OMSSA/NISTMS - Combining Spectral Library with Sequence Searching



Example from ~10,000 LTQ spectra, 1 fraction of a 2-D separation, breast tissue

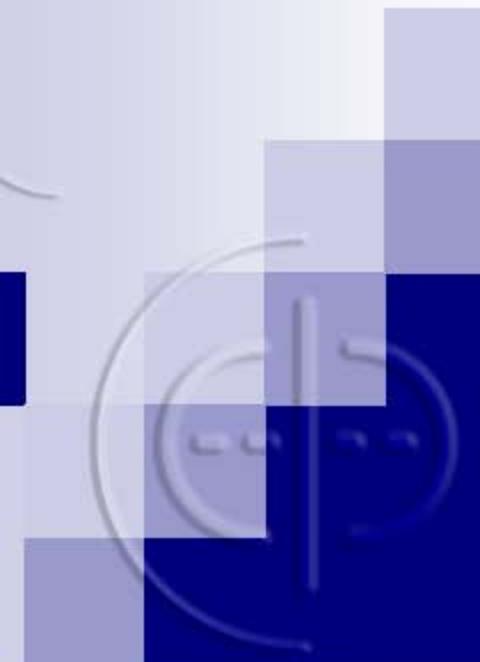
Pre-release NIST 2007 MS/MS Peptide Libraries coverage

Organism	Data Sources - Ion Trap			Library Spectra			Mean Spectra / Peptide
	LC-MS/MS Sample Runs (1D + 2D)	Total MS/MS Data files	*Total Peptide Identifications	Consensus Spectra	Increase (06-'07)	Peptide Sequences	
Human	274	56,469	17,779,962	179,857		4.1X	112,945
Yeast	59	3,044	**3,535,532	76,044		2.2X	46,819

* Search engines = *X! Tandem, OMSSA, MASCOT, ProteinProspector*; Threshold = 0.05 marginal FPR (F/I)

** Sequest IDs additionally included

- Modifications: ICAT, N-term acetyl, N-term pyro-glu, C Cam, M, Oxidation
- Includes ~10% semitryptic peptides and up to 2 missed cleavages
- Total amino acid coverage Human IPI (3.10): ~12%
- Total detectable amino acid coverage Human IPI (3.10): ~18%
 - digested with trypsin, 2 missed cleavages
 - m/z 300-2,000 Da



Protein Sequence Libraries

CLINICAL PROTEOMIC
TECHNOLOGIES FOR CANCER

Lewis Geer, NCBI/NLM/NIH

What would make a useful protein library for proteomics analysis?

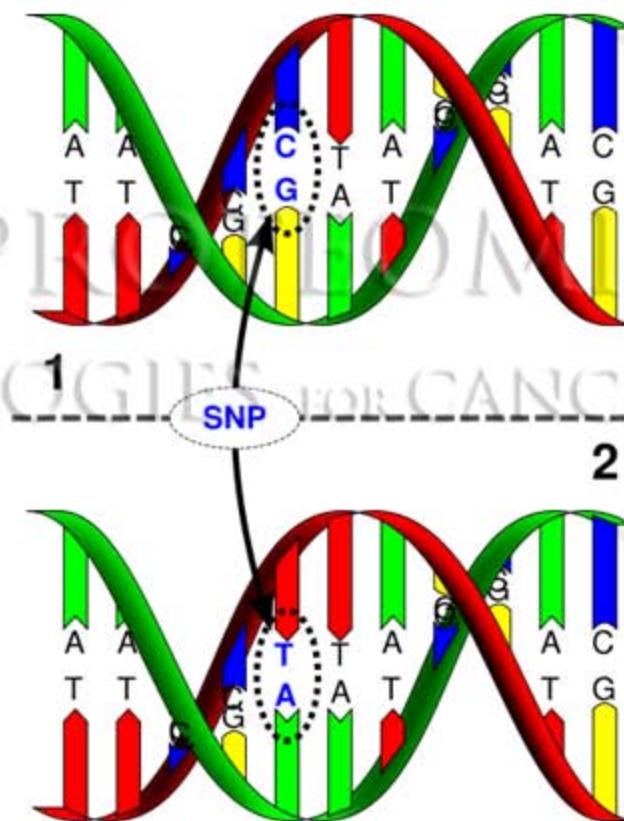
- Coverage: tradeoff with sensitivity

- cSNPs
 - Alternative splicing from cDNA evidence
 - Ab initio models
 - Mature peptides
 - Non canonical start/stop codons

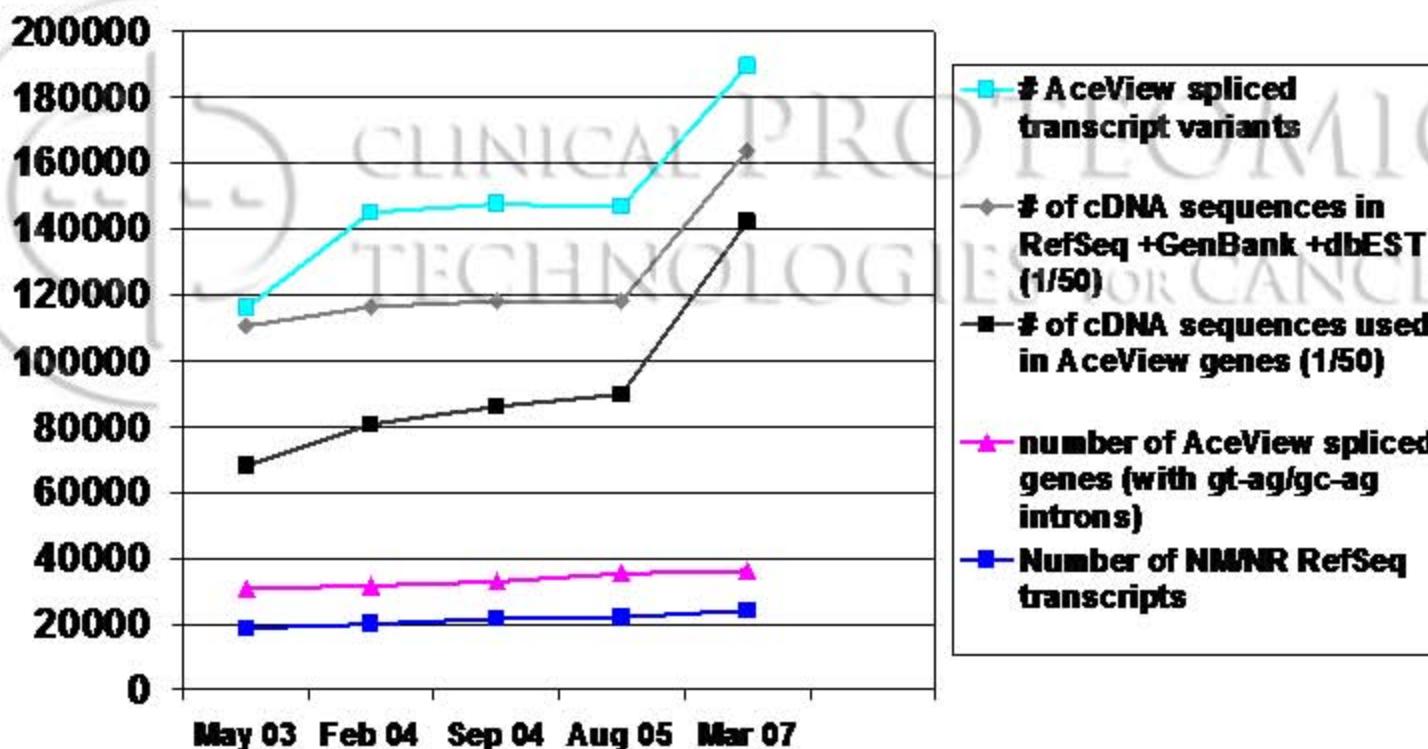
- Versioning

cSNPs

- SNP = single nucleotide polymorphism
- cSNP = coding SNP (also known as *non-synonymous SNP*).
- every 300 or so base pairs.



Alternative splicing from cDNA evidence



Ab Initio Prediction

Worm

400,000 cases

Human

A A G	GT A A G T T T T T	T T T T T T CAG A
A G	GT A A G T	T T T T T T CAG G
A A G	GCAA G T T T T T	T T T T T T T T CAG A
A G	GC A A G T	T T T T T T C AG G
A C	AT A T C C T T T	C AC AT

Mature peptides

LOCUS Q9NP84 129 aa linear PRI 02-OCT-2007
DEFINITION Tumor necrosis factor receptor superfamily member 12A precursor
(Fibroblast growth factor-inducible immediate-early response
protein 14) (FGF-inducible 14) (Tweak-receptor) (TweakR) (CD266
antigen).
[...] Region 1..27
/gene="TNFRSF12A"
/region_name="Signal"
/inference="non-experimental evidence, no additional
details recorded"
/note="Potential."
Region 28..129
/gene="TNFRSF12A"
/region_name="Mature chain"
/experiment="experimental evidence, no additional details
recorded"
/note="Tumor necrosis factor receptor superfamily member
12A. /FTId=PRO_0000034611."
[...] ORIGIN
1 margslrrl rllvlglwl llrsvageqa pgtapcsrgs swsadlldkcm dcascraph
61 sdfclgcaaa ppapfrllwp ilggalsltf vlglisgflv wrrcrrrekf ttpieetgge
121 gcpavaliq

Non canonical start/stop codons

LOCUS NM_001025366 3665 bp mRNA linear PRI 28-OCT-2007
DEFINITION Homo sapiens vascular endothelial growth factor A (VEGFA), transcript variant 1, mRNA.

[...]

```
misc_feature 645..647
/gene="VEGFA"
/note="Region: alternative non-AUG (CUG) translation
initiation site"
misc_feature 666..668
/gene="VEGFA"
/note="Region: alternative non-AUG (CUG) translation
initiation site"
misc_feature 906..908
/gene="VEGFA"
/note="Region: alternative non-AUG (CUG) translation
initiation site"
misc_feature 1032..1034
/gene="VEGFA"
/note="Region: alternative AUG translation initiation
site"
```

[...]

```
601 cggccggagg cgggtggag ggggtcgaaaa ctcgcggcgt cgcaactgaaa ctttcgccc
661 aacctctggg ctgttctcgc ttccggaggag ccgtggtccg cgcggggggaa gccgagccga
721 gcggagccgc gagaagtgtc agctcgggcc gggaggagcc gcagccggag gagggggagg
781 aggaagaaga gaaggaagag gagagggggc cgcagtggcg actcggcgt cggaagccgg
841 gctcatggac gggtgaggcg ggggtgtgcg cagacagtgc tccagccgcg cgcgtcccc
901 aggccctggc ccgggcctcg gggggggag gaagagtgc tcggccggagc gggaggaga
961 gggccggcc ccacagcccg agccggagag ggagcgcgag ccgcggccggc cccgtcgcc
1021 cttccgaaac catgaacttt ctgttgttt gggtgattt gggcatttcc ttgttgttct
```

[...]

Versioning

- What are the types of changes in version?
 - CrUD: Create, Update, and Delete
 - What does update mean?
 - change in annotation
 - change in sequence: what does “same sequence” mean?
- For synthetic records, CrUD can be due to algorithm features or changes in the algorithm
 - Genome builds
 - tracking of known genes
 - Clustered records
 - similarity calculation can be sensitive to algorithm changes and creates/deletes

Acknowledgements

■ NIST

- Stephen Stein
- Paul Rudnick
- and the rest of the NIST Chemical Reference Data Group

■ NCBI

- Steve Bryant
- Ming Xu
- Jean and Danielle Thierry-Mieg
- Kim Pruitt

■ NIMH

- Sanford Markey
- Douglas Slotta
- Jeffrey Kowalak
- rest of LNT

Calculating FDRs

Size of library Hits

$$FDR(o) = \frac{N(o,F)}{N(o,D)} * \frac{H(o,D)}{H(o,F)}$$

F = forward

D= decoy (assume all are false)

o=OMSSA

n=NISTMS

$$FDR(n) = \frac{N(n,F)}{N(n,D)} * \frac{H(n,D)}{H(n,F)}$$

$$FDR = \frac{[FDR(o) * H(o,F)] + [FDR(n) * H(n,F)]}{H(o,F) + H(n,F)}$$