Adventures at the proteome-genome boundary

DAVID L. TABB, PH.D.

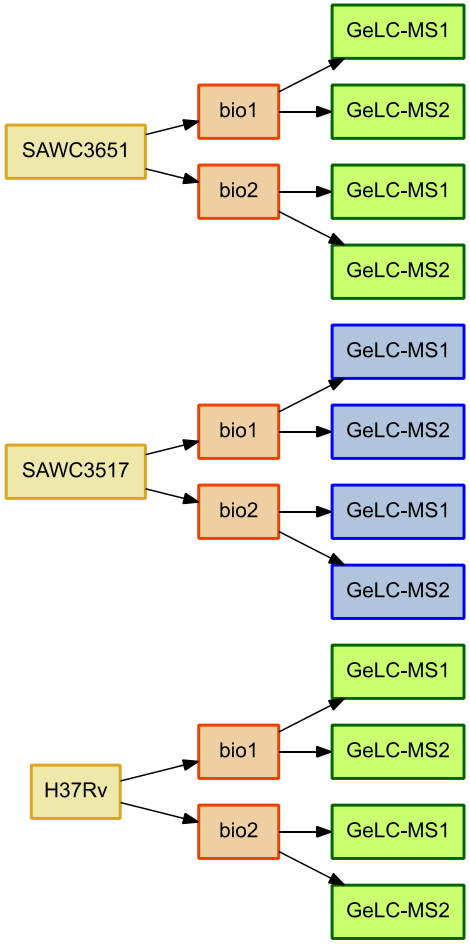
DIVISION OF MOLECULAR BIOLOGY AND HUMAN GENETICS
STELLENBOSCH UNIVERSITY
CAPE TOWN, SOUTH AFRICA

Outline

- "Classic" proteogenomics: expression of genomic variants and gene-finding
- Orthology alignment: recognition of evolutionary counterparts in paired proteomes
- Pre-genome proteomics: making proteomic DBs from RNA-Seq in non-model organisms

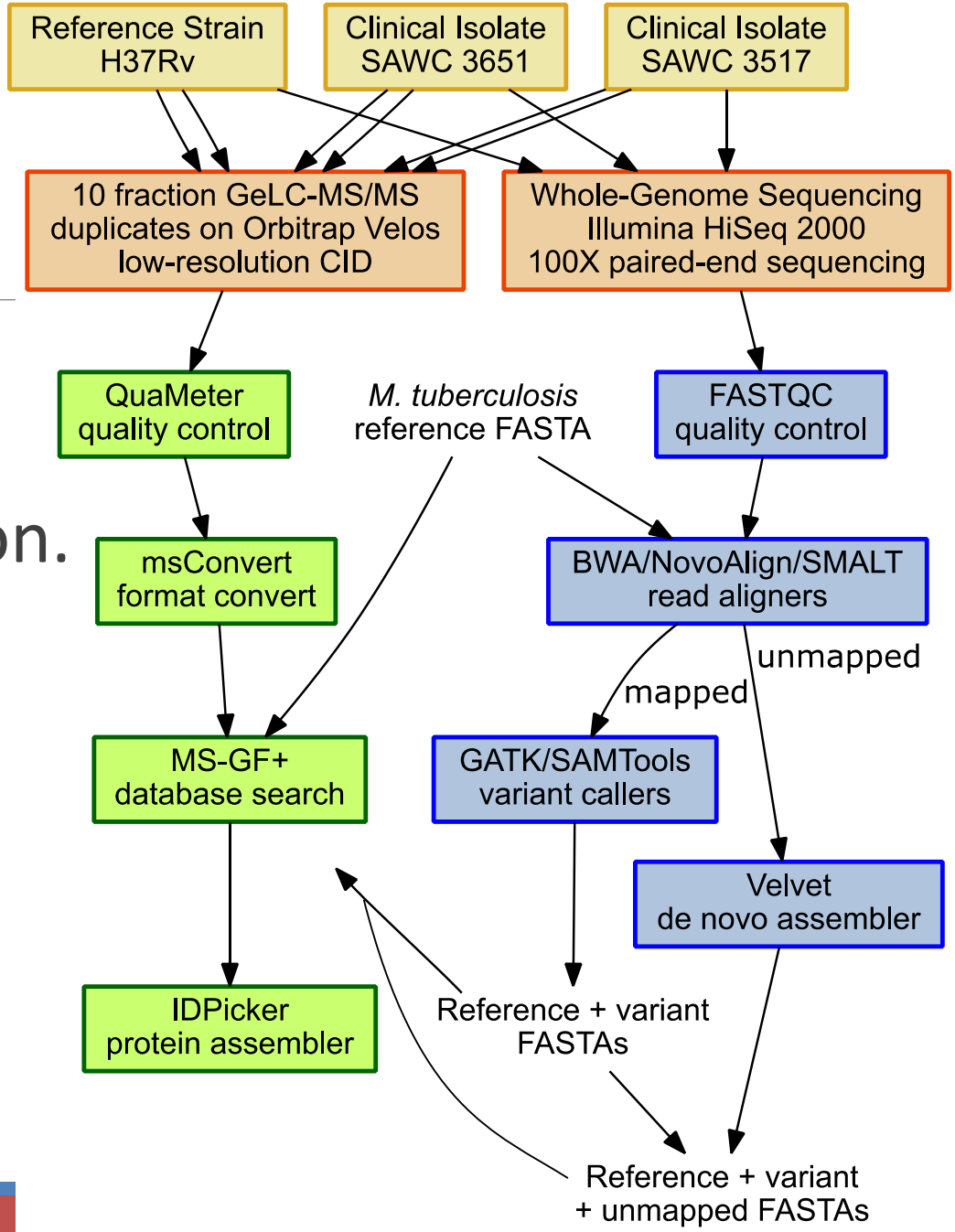
Proteogenomics: which genomic changes in *M.tb* strains are expressed?

- Two LAM *M.tb* strains were cultured from SA patients in the Western Cape.
- Genomic sequencing found major deletion separating them, with more variants separating both from the H37Rv reference strain.
- 10-fraction OrbiTrap Velos GeLC-MS inventoried peptides for six samples.



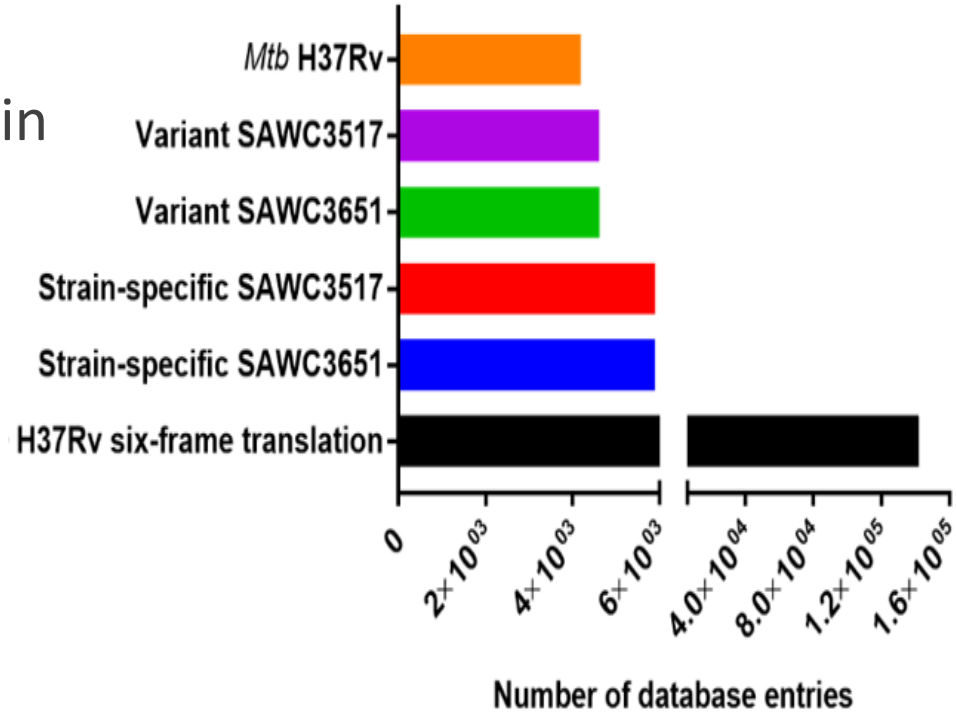
Data flow

- Both pathways receive *M.tb* reference annotation.
- Genome analysis informs proteome analysis, not vice versa.

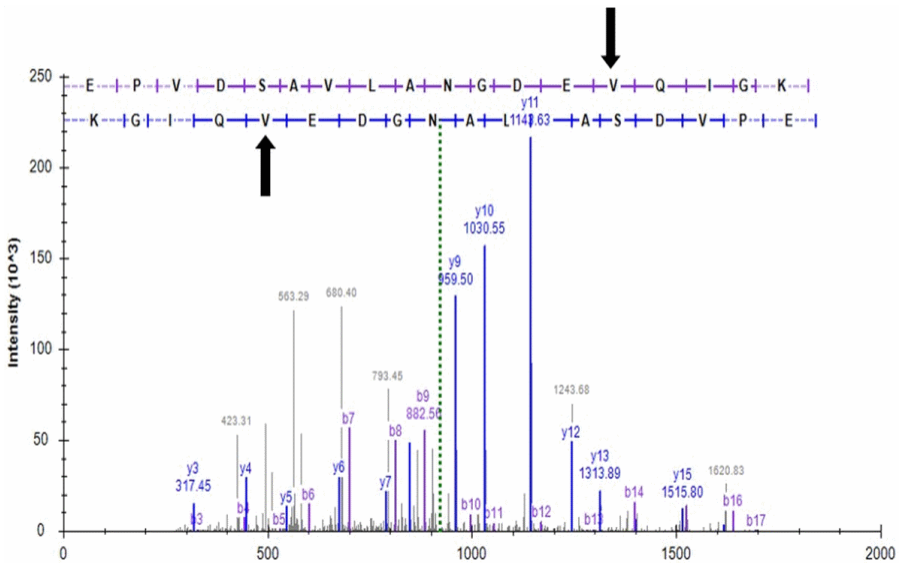


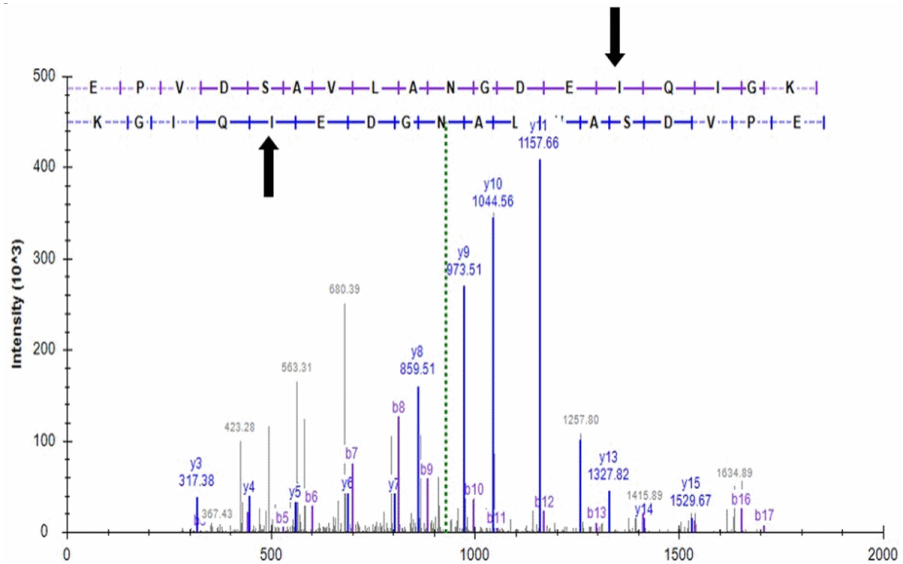
Adding strain-specific sequences impacts search space.

- Reference TubercuList proteome contains 4183 protein sequences.
- Applying nsSNVs to H37Rv made 422/430 additions.
- Adding unmapped regions via six-frame translation added ~1300 entries.
- "Six-frame" DB dwarfed the strain-specific databases.



Ion trap data are sufficient to detect variants

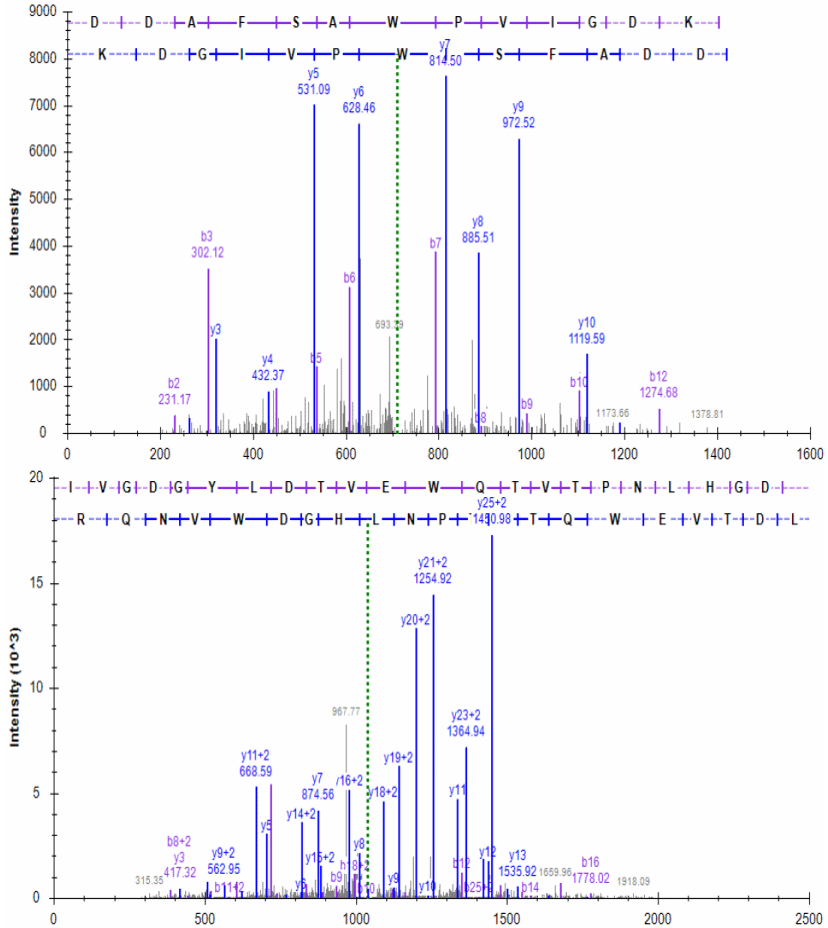




- Val becomes Ile at 137 in GarA for SAWC3651, changing both precursor mass and most of the y" ion masses.
- We identified a total of 72 variant peptides, with 59 passing manual inspection of MS/MS evidence.

Interpreting the hits to unmapped DNA

- Some peptides hitting "novel" genes also matched reference sequences.
- Peptides matching deleted regions could be explained by undeleted paralogs.
- 29 peptides supported novel helicase from SAWC3651.



Spectra from novel helicase

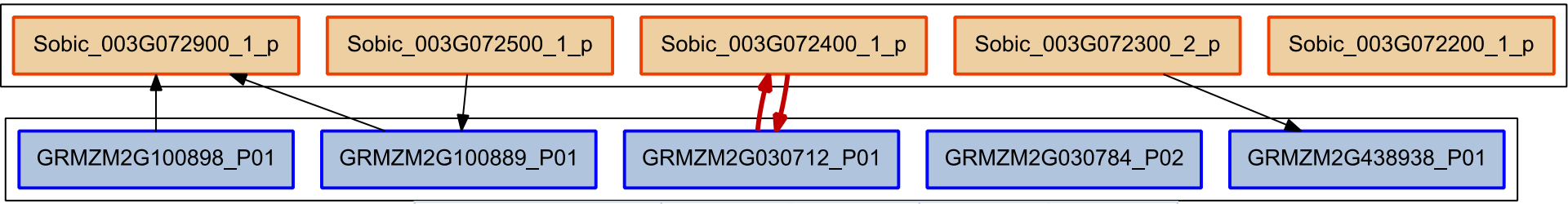
Orthology alignment: linking proteomes of two species

- Genomic information can facilitate comparison between species.
- UWC seeks proteins changing disparately between sorghum and maize in response to water availability.

	Drought-like	Well-watered
	5 replicates	5 replicates
	5 replicates	5 replicates

Orthology detection and application

- BLAST compares all proteins of *Sorghum bicolor* (47k) and *Zea mays* (89k).
- Scripts align ortholog maize and sorghum protein expressions to same row.



Original	Sorghum		Maize	
	Wet	Dry	Wet	Dry
Sobic_003G072400_1_p	A B C D E	F G H I J	X X X X X	X X X X X
GRMZM2G030712_P01	X X X X X	X X X X X	K L M N O	P Q R S T
Aligned	Sorghum		Maize	
	Wet	Dry	Wet	Dry
Both	A B C D E	F G H I J	K L M N O	P Q R S T

Why align orthologs?

- Annotation in plant proteomes is relatively immature; we may know function of gene in one species but not in another.
- Statistical models can find differential proteins more robustly with twice the data.
- Alignment more than *doubled* the number of rows that described both maize and sorghum orthologs!

Our biological question cannot be answered from two disjoint sets.

Chia: from RNA-Seq to annotated proteome

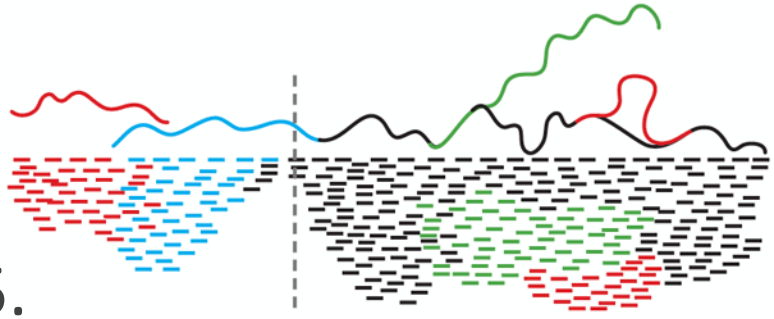
- UWC conducts proteome research in chia, but they are hindered since UniProt contains only 9 sequences for *Salvia hispanica*.
- Sreedhar *et al* published RNA-Seq data from an Illumina GAIIx for five samples.
- Could we build a more complete list of proteins for this species by *de novo* assembly?

essenceofthedesert.wordpress.com

R.V. Sreedhar et al. *PLOS One* (2014) 10(4): e0123580

Trinity *de novo* assembly

- Sreedhar *et al* reported 76,014 transcripts from Trinity, but in our hands the software produced 55,465.



- We chose two routes to annotate our transcripts:
  - tBLASTn against the nearest annotated species
  - InterPro to recognize motifs seen in other species

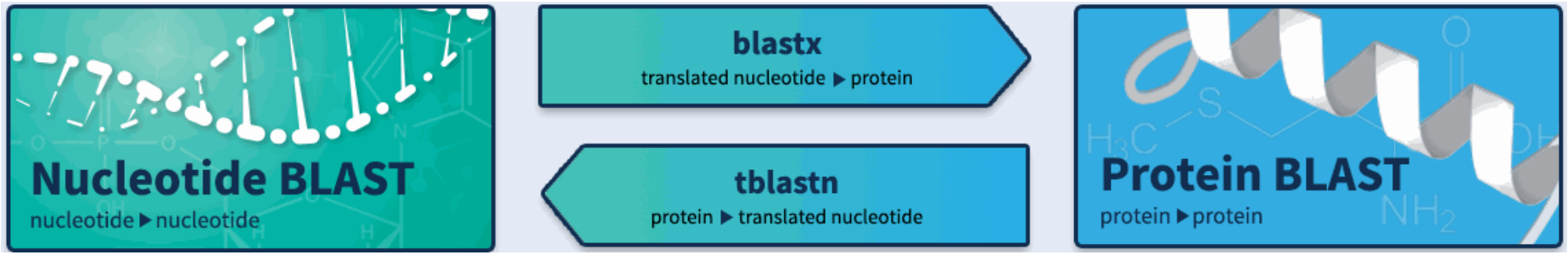


Image from Iyer and Chinnaiyan, *Nat. Biotech*. (2011) 29: 599-600

Our shortcut points to four complete annotations

- *Paraboea paniculata*
  - an African violet
  - 50,113 proteins

- *Genlisea aurea*
  - carnivorous with compact genome
  - 17,693 proteins

- *Sesamum indicum*
  - "Sesame"
  - 33,467 proteins

- *Erythranthe guttata*
  - spotted monkey flower
  - 61,983 proteins

murvegetalpatrickblanc.com, cpukforum.com, wikimedia.org, nzflora.info

UNIVERSITEIT STELLENBOSCH UNIVERSITY

What works in chia may work for the spotted hyena

- SUN researchers in wildlife TB see *Crocuta crocuta* successfully resisting infectious disease.
- UniProt yields only 162 proteins.
- We have conducted RNA-Seq on an Illumina NextSeq-500 at CPGR, using four different tissues.
- After assembly, we will refer to closest relatives: mongooses and meerkats, then cats


Yathin S Krishnappa

Work with SDC Parsons and MA Miller, featuring Roxanne Higgitt

Takeaway Messages

- Proteogenomics can take many forms because each has much to offer the other.
- Bioinformatics frequently requires side-trips into biostatistics to offer its full value.
- “Non-model” organisms can rapidly acquire annotation via inexpensive sequencing.

Acknowledgment

- SUN Division of Molecular Biology and Human Genetics
- UWC Department of Biotechnology
- Centre for Proteomic and Genomic Research
- SA Medical Research Council
  "South Africa Tuberculosis Bioinformatics Initiative"