Finding “missing” proteins - sniffing out olfactory receptors

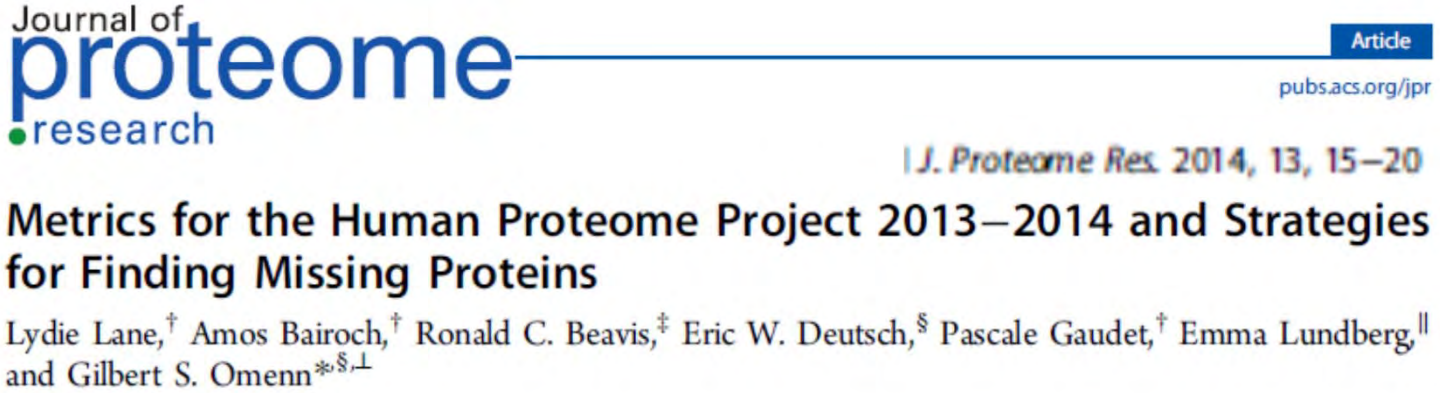
Shoba Ranganathan

Abidali Mohamedali

Ishmam Nawar

So what are “missing” proteins?

- Of the 20 123 Ensembl protein encoding genes (Sept. 13), NextProt has high quality evidence for 15 646 proteins
  - credible evidence of protein expression and identification by mass spectrometry, immunohistochemistry, 3D structure, or amino acid sequencing
  - 3844 proteins currently classified as “missing” after discounting 638 dubious genes or pseudogenes.



Evidence for missing proteins

- Proteomics Evidence
- Transcript Evidence
- Antibody Evidence
- Disease Evidence
- 3D Structural Evidence

NeXtProt- CHPP Repository currently identifies **3954** Proteins with at least one “no” in each category

Missing Proteins as per NeXtProt

	Major (>10) Missing protein families	number
1	ring finger proteins	12
2	taste receptors	16
3	potassium channel proteins	18
4	late cornified envelope proteins	19
5	nuclear proteins	24
6	defensins	28
7	ankyrin repeat & death-domain proteins	39
8	other GPCRs	40
9	homeobox proteins	41
10	cadherins	49
11	leucine-rich repeat proteins	51
12	keratin-associated proteins	61
13	coiled-coil domain proteins	65
14	transmembrane proteins	117
15	zinc finger proteins	308
16	**olfactory receptors**	**439**
		1327

Recent “draft human proteomes” and olfactory receptors (ORs)

- The Pandey paper did not sequence any nasal/mouth/throat/tongue epithelia and so cannot be expected to report any ORs.
- So there is low probability of proteomic evidence for ORs systematically uncovered here.
- So far, the Pandey data provides MS evidence for 1 known OR (Q8NGA1) and 1 “missing” OR (Q9GZK7).
- So to sniff out ORs, can we use Kuster’s approach?
- Where to start?

How many ORs are currently "missing"?

- NextProt lists 439 ORs as missing.
- Of these, some have MS (26) or antibody evidence (109) or both (4).
- Also these are numerous redundancies as the same protein sequence can be encoded by several transcription start sites on the genome.
- So, we have first worked on removing redundancies and identifying sequences with some proteomics and/or antibody evidence.

NR OR list by Chromosome

"nr-really-missing" – no MS, antibody or 3D structure evidence.
"nr-partly-missing" – some evidence – not to NeXtProt standards

Grand total = 389
(50 redundancies:

- 48 in Chr6,
- 1 in Chr 1 and
- 1 in Chr7)

Chr	nr-really-missing	nr-partly-missing
1	42	19
2		2
3	10	1
5	1	3
6	8	8
7	10	5
8		1
9	16	7
10		1
11	122	45
12	12	5
14	18	7
15	3	2
16	2	
17	13	
19	15	4
22		1
X	1	
unk	4	1
	277	112

Hybrid approach

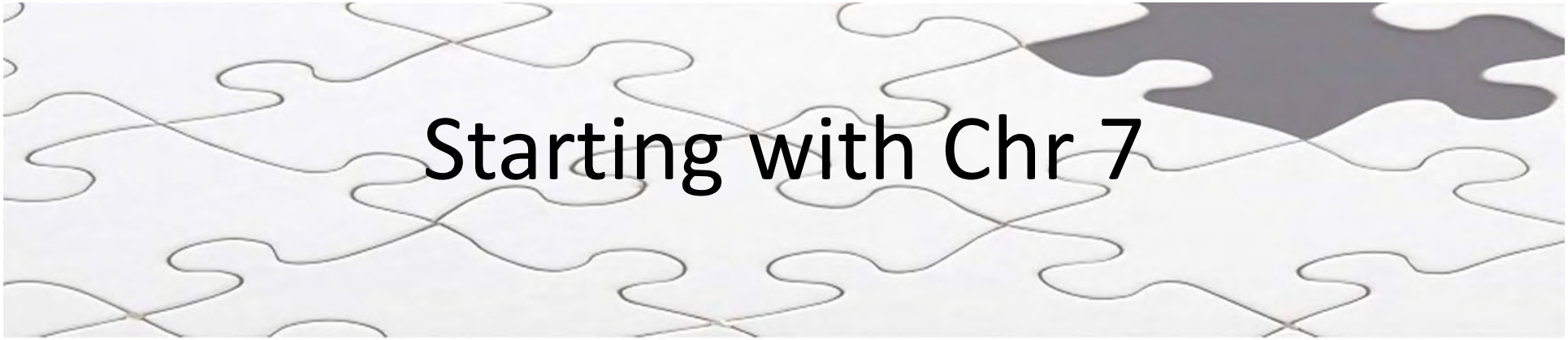
- Kuster's approach is good: there is evidence out there
- However, it must be good quality.
- "One peptide does not a protein make!"
- Unique peptides are more credible than ones mapping to many proteins albeit of the same functional class.

Missing OR Proteins per Chromosome

Chromosome	
1	61
2	2
3	11
5	4
6	16
7	15
8	1
9	23
10	1
11	167
12	17
14	25
15	5
16	2
17	13
19	19
22	1
X	1
unk	5
	389

Checked against

- PRIDE DB
- GPMDB
- HUMAN PROTEINPEADIA
- PROTEOMICS DB (KUSTER)



Starting with Chr 7

- There are 15 missing ORs (+1 redundancy).
- Where should we look for evidence?
- Proteomic databases:
  1. PRIDE
  2. GPMDB (from PeptideAtlas)
  3. Proteinpedia (from HPRD – the old Pandey database)
  4. Proteomics DB (Kuster)
- How do we know we can count on this evidence?
  - BLASTP peptide (if any) against human ref sequences
    - If unique, we count that evidence.
    - If not, we use it as supporting evidence.

Evidence Found!

Example:

OR9A2- HUMAN Olfactory receptor 9A2

- 4 proteotypic peptides seen in PRIDE
- 1 proteotypic peptide seen in Proteomics DB
- Additional evidence in literature
- Most with good Sequest scores (>200)
- In 6 different experiments
- In PLASMA, HUVEC Cells and Hepatocyte cells

>sp|Q8NGT5|OR9A2_HUMAN Olfactory receptor 9A2 OS=Homo sapiens GN=OR9A2 PE=2 SV=1

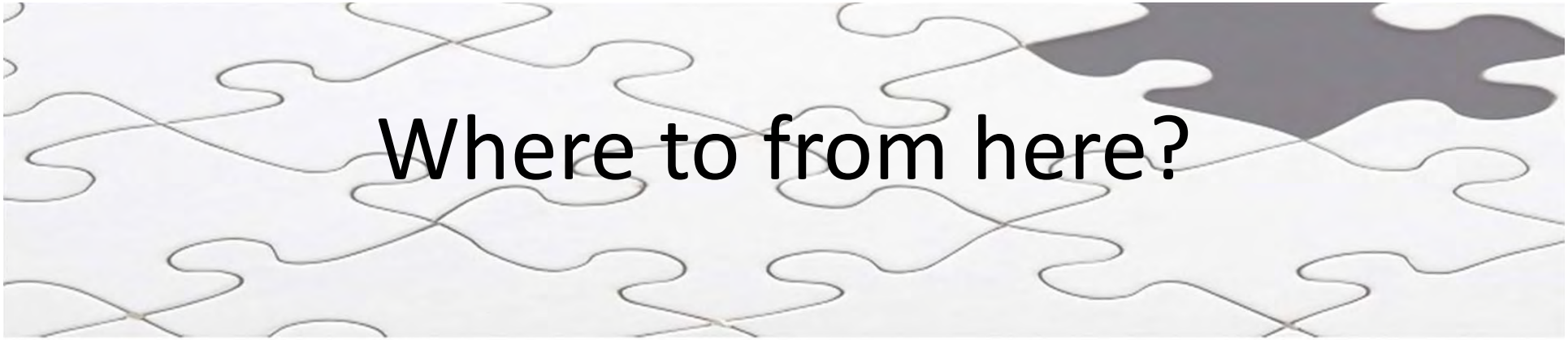
MMDNHSSATEFHLLGFPGSQGLHHILFAIFFFYLVTLMGNTVIIVIVCVDKRLQSPMYF
FLSHLSTLEILVTTIIVPMMLWGLLFLGCRQYLSLHVSLNFSCGTMEFALLGVMAVDRYV
AVCNPLRYNIIMNSSTCIWVVIVSWVFGFLSEIWPIYATFQFTFRKSNSLDHFYCDRGQL
LKLSCDNTLLTEFILFLMAVFILIGSLIPTIVSYTYIISTILKIPSASGRRKAFSTFASH
FTCVVIGYGSCLFLYVKPKQTQGVEYNKIVSLLVSVLTPFLNPFIFTLRNDKVKEALRDG
MKRCCQLLKD

KEY:

PRIDE

Proteomics DB

Coverage: 25%



Where to from here?

- Interrogate all possible databases for any evidence already present but not easily accessible
- Text mining of other forms of evidence (based on synonyms) from literature
- Determine what proteins are truly ‘missing’

Current work on Chr7 “missing” protein list

- ORs
- Taste receptors
- Other GPCRs

We are going to sniff them out!