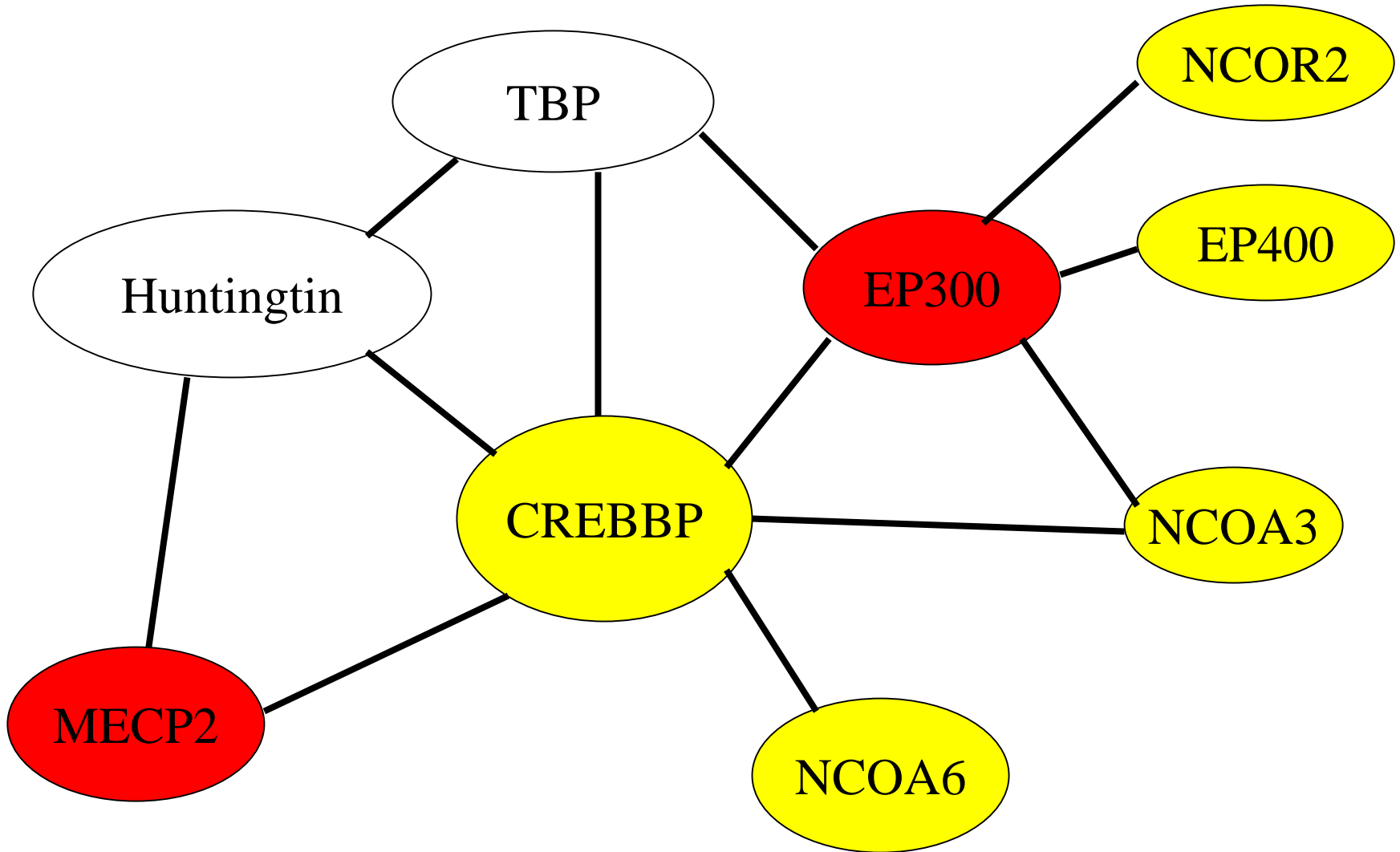


Common evidence network: Investigating Medline co-citations of candidate disease genes

Alison Meynert – December 15, 2004
Supervisors: Francis Ouellette (UBiC)
and Anne Condon (UBC Computer Science)

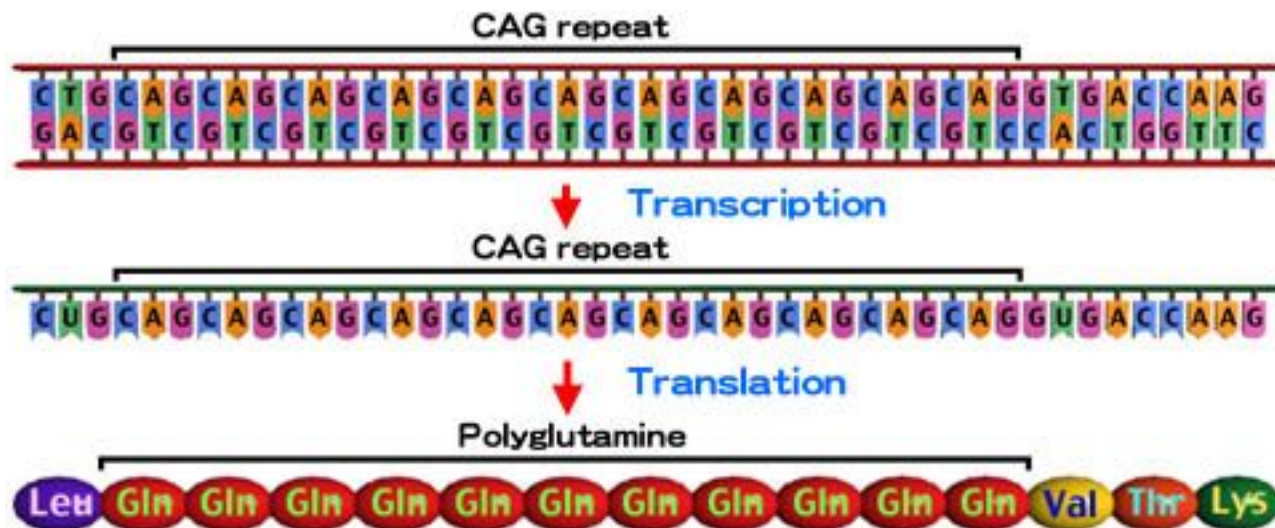


What kind of evidence links genes?

- Medline co-citation
- Entry in a protein-protein interaction database
- Part of the same pathway (i.e. KEGG)
- Similar co-expression patterns
- Shared overrepresentation of Gene Ontology terms

Which genes are we interested in?

- GeMS: Genomic Mutational Signatures Project
 - Identification of novel disease-related genes via signature sequence mutations
 - Focus: genes encoding poly-glutamine tracts



Medline XML citation example

- <MedlineCitation Owner="NLM" Status="Completed">

```

+ <PMID>- <MeshHeading> - <Abstract> - <ChemicalList>
+ <DateCreated> - <MeshHeading> - <AbstractText> - <ChemicalList> - <DataBankList CompleteYN="Y">
+ <DateCompleted> <DescriptorName> Hunting - <ArticleTitle>
  
```

(BDNF; ref. 2). Here we show that the repressed wild-type huntingtin activity on BDNF promoter

is the target of silencing activity

```

+ <PageNumber>1125</PageNumber>
+ <AbstractText>
+ <DataBankList CompleteYN="Y">
+ <Affiliation>
+ <AuthorList AuthorName="Huntington M, et al.">
+ <Language>
+ <PublicationTypeList>
+ <ElectronMicroscopy>
+ <MedlineCitation Owner="NLM" Status="Completed">
+ <ChemicalList>
+ <CitationList>
+ <Comments>
+ <GeneSymbol>
+ <MeshHeadingList>
+ <Keywords>
</MedlineCitation>
  
```

increasing cytoplasmic REST/NRSE-cofactor activity in neurons to regulate the availability of REST/NRSE to its nuclear NRSE-binding site and that this control is lost in the pathology of Huntington disease. These data identify a new mechanism by which mutation of huntingtin causes loss of transcription of neuronal genes.

How do we identify citations of interest?

1. For genes of interest, build a list of gene names and synonyms, including accession numbers, from NCBI LocusLink/Gene
2. Parse Medline XML files and identify sections of interest (i.e. abstract, gene symbols)
3. Break text into words where required
4. Look for matches to gene synonyms
5. Output the PMID, matched synonym, primary gene symbol, and name of XML element
6. Results are loaded into a MySQL database

What about false positive matches?

- Many gene names/symbols are ambiguous
 - Huntingtin: HD = Hodgkin Disease
 - Androgen Receptor: KD = disassociation constant
- Main method: examine MeSH headings associated with citations where an ambiguous gene synonym has been matched
 - Some headings are immediately indicative of a false positive match





Pruning the MeSH ontology tree

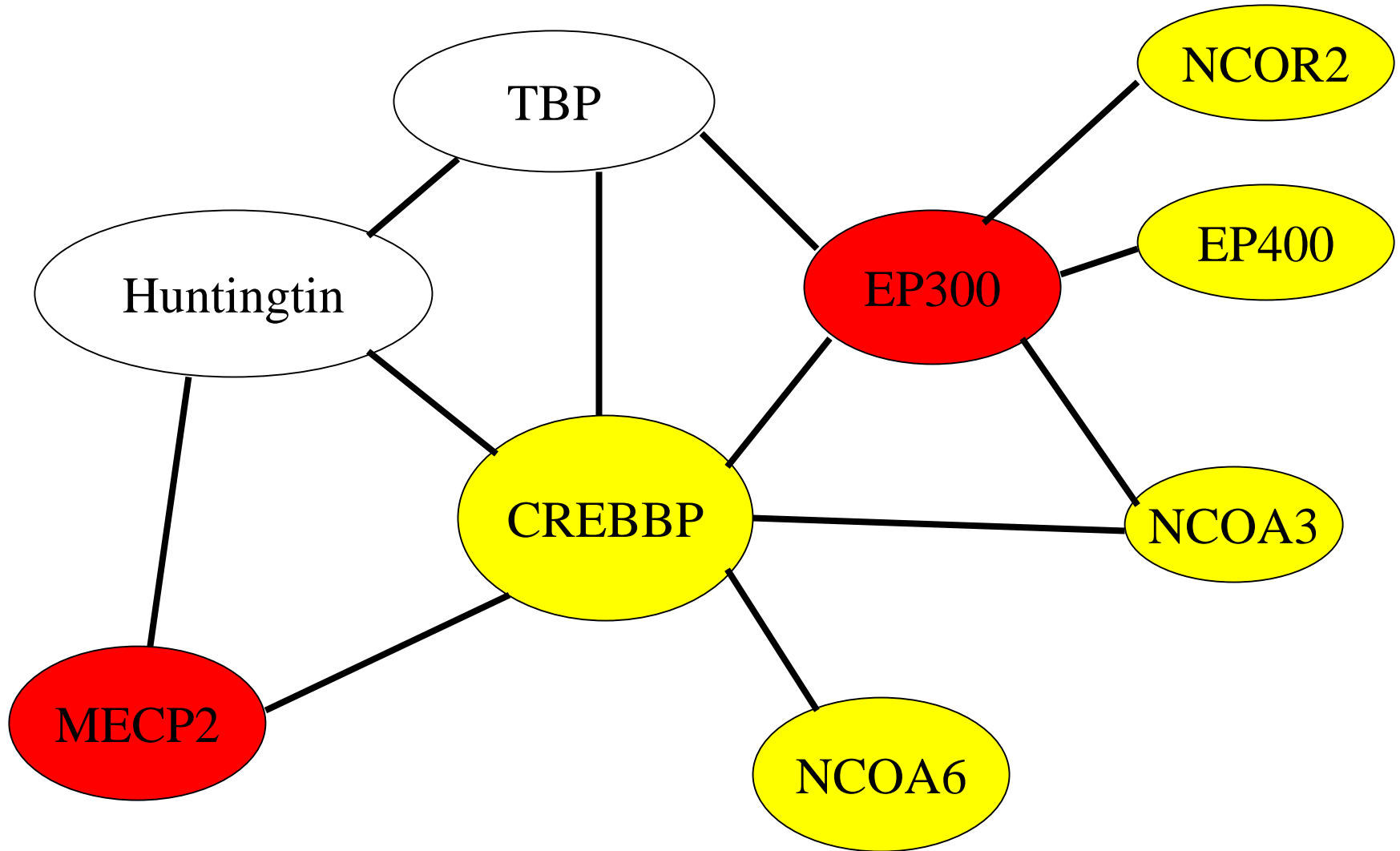
- + Anatomy [A]
 - + Organisms [B]
 - + Diseases [C]
 - + Chemicals and Drugs [D]
 - + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
 - + Psychiatry and Psychology [F]
 - + Biological Sciences [G]
 - + Physical Sciences [H]
 - + Anthropology, Education, Sociology and Social Phenomena [I]
 - + Technology and Food and Beverages [J]
 - + Humanities [K]
 - + Information Science [L]
 - + Persons [M]
 - + Geographic Locations [Z]
-
- ◊ [Health Care Economics and Organizations \[N03\]](#) +
 - ◊ [Health Services Administration \[N04\]](#) +
 - ◊ [Health Care Quality, Access, and Evaluation \[N05\]](#) +
- + Geographic Locations [Z]

Current results

Match count	Not false positives	XML location of match	Method of analysis
113,706	93,031	Abstract	Automated, filter by associated MeSH headings
12,475	7,818	Title	
5,537	4,719	Chemical	
484	401	Gene symbol	
12,267	6,355	Mesh heading	Manually verified
88	14	Keyword	
365	365	Accession	Trusted as true positives
144,922	112,703	All locations	

Next steps/further work

- False positive filtering:
 - Finish filtering based on MeSH heading association with ambiguous synonym matches
 - Add filtering based on phrases that are not associated with a particular MeSH heading
 - If required, additional natural language processing
- Run synonym matching code on all genes in NCBI LocusLink/Gene on our citation dataset
- Provide weighting for different evidence sources
- Visualize and analyze co-citation network



Acknowledgements

Supervisors

- Anne Condon
- Francis Ouellette

Labs

- UBiC/CMMT Bioinformatics Lab
- UBC Beta Lab

GeMS Project

- Stefanie Butland
- Yong Huang
- Soo Sen Lee
- George Yang
- Blair Leavitt
- Carri-Lyn Mead

Alison's research is supported by NSERC, Simon Fraser University, and the CIHR/Michael Smith Foundation Bioinformatics Training Program for Health Research.