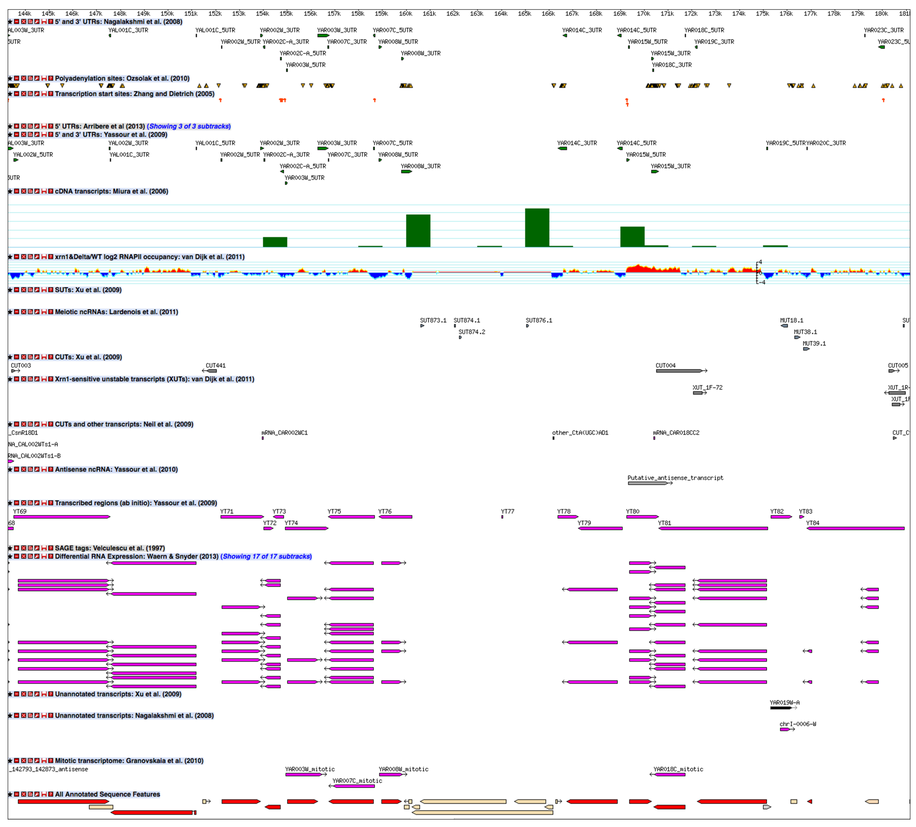
Defining the transcriptome of *Saccharomyces cerevisiae*

Janos Demeter, Paul Lloyd, Edith D. Wong, J. Michael Cherry
Saccharomyces Genome Database, Department of Genetics, Stanford University

The transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in a single cell or a population of cells. Defining the complete transcriptome of the budding yeast, a single celled eukaryotic organism, should be an achievable goal with the use of high-throughput genomic technologies (e.g. tiling microarrays, next generation sequencing). Indeed, numerous publications have addressed this question and made their datasets publicly available. As a scientific database that provides researchers with high-quality curated data, the *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) set out to collect a representative collection of these high quality and frequently cited datasets with the goal of compiling them into a complete transcriptome in yeast - a longstanding request of our users. Integration of these datasets was more complicated than anticipated for a number of reasons. First, exact matches between the datasets did not even remotely cover the genome completely. One major issue is the various technologies producing different kinds of results. E.g.: some approaches give only the 5'-end, others only the 3'-end and yet others both ends of the messages. Our integrative approach raises the question whether to define transcription start and end points or entire transcripts. Another issue to be resolved is the detection of low abundance transcripts and their functional significance. We present the results of our first attempt to define a transcriptome for *S. cerevisiae.* Initially, we confined ourselves to a single common condition, wild type lab strains grown in rich medium. In the future, we will extend the analysis to a much wider list of conditions.

INTRODUCTION



Goal: To compile a dataset that represent the *S. cerevisiae* transcriptome.

Many datasets that address this question on a genomic scale.

paper	technique	target	pmid
Lardenois A, et al. (2011)	tiling arrays	ncRNA	21149693
Miura F, et al. (2006)	cDNA clone lib	TSS	17101987
Nagalakshmi U, et al. (2008)	RNA-Seq (single end)	transcripts	18451266
Neil H, et al. (2009)	3' sage/tiling arrays	CUTs	19169244
Ozsolak F, et al. (2010)	direct rna sequencing	polyA-sites	21145465
Pelechano et al. (2013)	TIF-Seq	paired TSS-TTS	23615609
van Dijk EL, et al. (2011)	RNA-Seq (single end)	XUTs	21697827
Xu Z, et al. (2009)	tiling arrays	transcripts	19169243
Yassour M, et al. (2009)	RNA-seq (single end)	transcripts	19208812
Yassour M, et al. (2010)	stranded RNA-Seq (paired ends)	antisense	20796282
Zhang Z and Dietrich FS (2005)	5' sage	TSS	15905473

Issues:

- What is the transcriptome? (“the set of all RNA molecules … produced in one or a population of cells.”)
- Should a known function be part of the definition? What about “pervasive transcription”?
- Practical issues:
  - Various technologies
  - Studies have different aims
  - No definitive dataset
  - Limited set of conditions
  - Pelechano et al dataset is much larger than others

DATA INTEGRATION

In this part of the project we looked at only pre-analyzed datasets.

Possible approaches:

- Find exact matches across datasets
- **Find matches to Pelechano et al across all datasets**
- Use Pelechano et. al. as the most recent and largest dataset

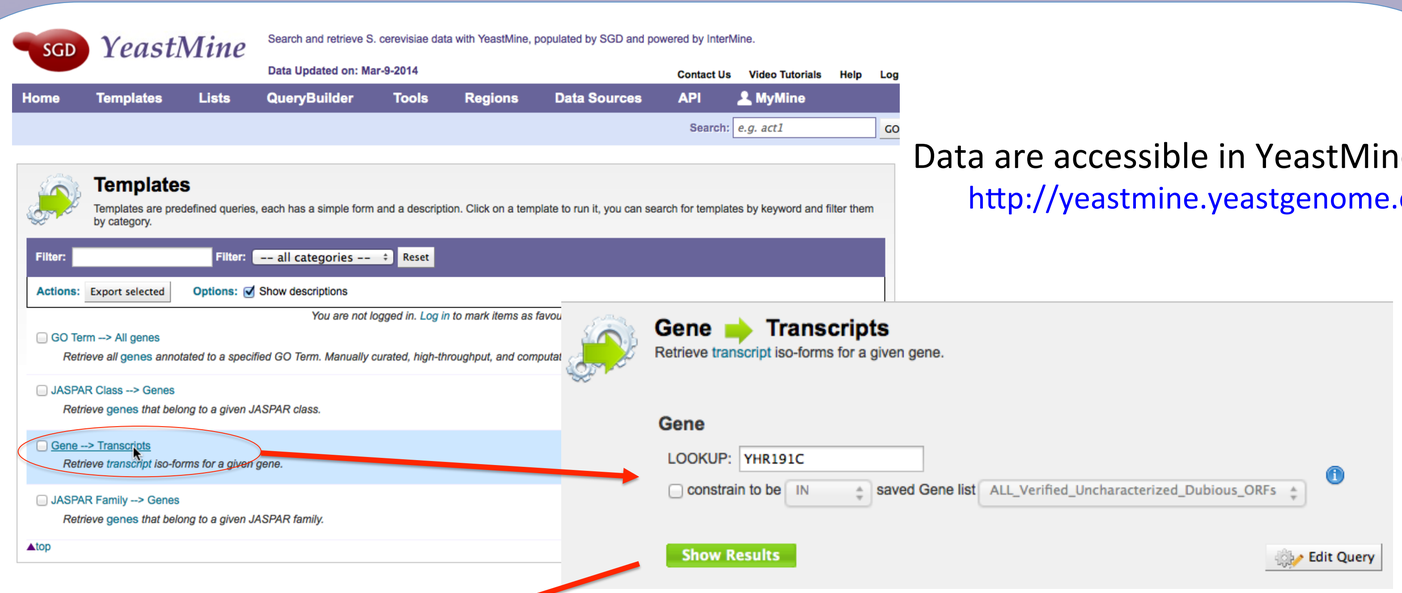
Pelechano et al mTIF groups	count
mTIFs	371087
mTIFs covering single ORFs (aa)	184473
mTIFs overlaping 3' of one ORF	135806
mTIFs intergenic transcripts	16824
mTIFs overlaping 5' of one ORF	16761
mTIFs covering >=2 ORFs	4140
mTIFs overlaping >=2 ORFs	4096
mTIFs internal transcripts	3681
mTIFs for SUT/CUT/XUT	5306
no of ORFs (aa)	5141
no of verified ORFs (aa)	4492
no of uncharacterized ORFs (aa)	480
no of dubious ORFs (aa)	169
unique 5' ends (aa)	49196
unique 3' ends (aa)	49424

- Treat 5’ and 3’ ends separately
- Allow imperfect matches (window +/- 3 nt)
- Score by counting datasets with matching coordinates
- Keep mTIFs with score >= 1
- Keep fully covered ORFs only

SGD transcriptome	count
mTIFs	22623
mTIFs covering single ORFs (aa)	17906
mTIFs overlaping 3' of one ORF	588
mTIFs intergenic transcripts	1153
mTIFs overlaping 5' of one ORF	186
mTIFs covering >=2 ORFs	1398
mTIFs overlaping >=2 ORFs	120
mTIFs internal transcripts	2
mTIFs for SUT/CUT/XUT	1270
no of ORFs (aa)	2850
no of verified ORFs (aa)	2666
no of uncharacterized ORFs (aa)	181
no of dubious ORFs (aa)	3
unique 5' ends (aa)	8313
unique 3' ends (aa)	6067

mTIF: major transcript isoform

DATA AVAILABILITY



Data are accessible in YeastMine:
http://yeastmine.yeastgenome.org

Showing 1 to 10 of 10 rows

Systematic Name	Transcripts > Primary DBID	Transcripts > Standard Name	Chromosome > Identifier	Chromosome Location > Start	Chromosome Location > End	Transcripts > Glucose Count	Transcripts > Galactose Count	Transcripts > Note	Transcripts > Five Prime Score	Transcripts > Three Prime Score	Transcripts > Five Prime Data Set	Transcripts > Three Prime Data Set
YHR191C	S000157711	SC_Transcript_00000733	chrVIII	486048	486691	1	1	Covering_one_intact_ORF	1	1	Xu_2009_ORFs	Nagalakshmi_2008
YHR191C	S000158863	SC_Transcript_00001885	chrVIII	486202	486672	0	3	Covering_one_intact_ORF	1	1	Nagalakshmi_2008	Miura_2006
YHR191C	S000162482	SC_Transcript_00005504	chrVIII	486074	486672	6	12	Covering_one_intact_ORF	1	1	Nagalakshmi_2008	Xu_2009_ORFs
YHR191C	S000162629	SC_Transcript_00005651	chrVIII	486202	486667	22	32	Covering_one_intact_ORF	1	1	Miura_2006	Miura_2006
YHR191C	S000164409	SC_Transcript_00007431	chrVIII	486074	486667	41	54	Covering_one_intact_ORF	1	1	Miura_2006	Xu_2009_ORFs
YHR191C	S000165424	SC_Transcript_00008446	chrVIII	486074	486691	12	9	Covering_one_intact_ORF	1	1	Xu_2009_ORFs	Xu_2009_ORFs
YHR191C	S000168837	SC_Transcript_00011859	chrVIII	486069	486672	2	0	Covering_one_intact_ORF	1	1	Nagalakshmi_2008	Xu_2009_ORFs
YHR191C	S000169700	SC_Transcript_00012722	chrVIII	486048	486667	3	0	Covering_one_intact_ORF	1	1	Miura_2006	Nagalakshmi_2008
YHR191C	S000169845	SC_Transcript_00012867	chrVIII	486202	486691	8	17	Covering_one_intact_ORF	1	1	Xu_2009_ORFs	Miura_2006
YHR191C	S000171838	SC_Transcript_00014860	chrVIII	486069	486667	9	6	Covering_one_intact_ORF	1	1	Miura_2006	Xu_2009_ORFs

FUTURE PLANS

To improve coverage:

- Include additional datasets
- Expand the conditions covered beyond rich media
- Add datasets with raw data only by processing fastq files
- Incorporate results from small scale studies

Leverage resulting transcriptome data for other SGD projects