Estimation of sequence errors and prediction capacity in transcriptomic and DNA-protein interaction assays

Eric Rivals

LIRMM - Méthodes Algorithmes pour la Bioinfo

www.lirmm.fr/~rivals

Transcriptomics

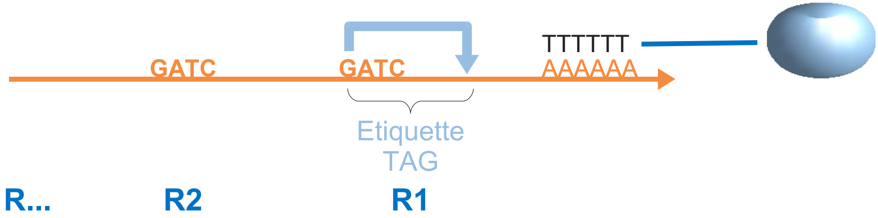
Transcriptome: all RNAs present in a cell

- Transcriptomics: identify and count each RNA of a cell
  sequence and genomic region of origin
- Techniques:

by sequencing	: EST, SAGE, MPSS, CAGE, etc.	open
by hybridisation	: DNA arrays	close
	"Whole" Genome Tiling Array	open

- Which diversity of transcripts in a cell?
- 70% of human or mouse genome is transcribed
  RNA dark matter [Zarmore, Science, 05]
- Which genomic regions are transcribed? in which conditions?

Serial Analysis of Gene Expression SAGE [Velculescu et al. 95]



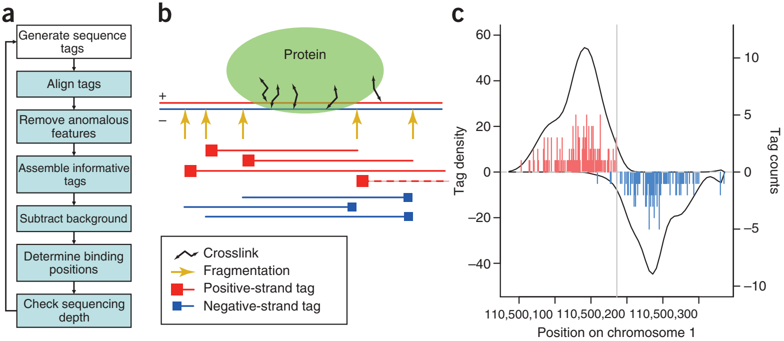
- anchor site 4 pb: usually CATG with NlaIII enzyme
- tags are 14 (SAGE) or 21 pb long (LongSAGE)
- *occurrence*: number of copies observed for a given transcript

Sequence census assay

Chromatin ImmunoPrecipitation with sequencing (ChIP-seq)

ChIP-seq is a method to identify genome-wide DNA binding sites for a protein of interest

E.g., polymerase, transcription factors, histone modification, etc.



[Kharchenko et al., Nat. Biotech., 08]

Next generation sequencing technologies

- Sequencing in paralell millions of short sequence reads
- on a single apparatus, in a few hours
- Technologies: 454®(pyro-sequencing), Illumina®/Solexa
- Examples:

PMAGE: PCR-colonies, sequencing by ligation, [Kim et al., 07]
2.3 million of 14 bp occurrences for 72 K tags

SAGE-Solexa: combines LongSAGE with Solexa
2.2 million occurrences for 440 445 tags of 21 bp

ChIP-seq: combines Chromatin ImmunoPrecipitation with Solexa
1.5 million of 25 bp sequences [Johnson et al. 07]
15 millions of 20 bp reads [Boyle et al. 08]

Mapping

- Find for each tag all genomic positions at which the tag match either exactly or approximately on the human genome (+/− strands)
- Fast exact mapping simultaneously for large tags sets with MPSCAN [Rivals et al., submitted]
- Results: is a tag located? once or more than once?

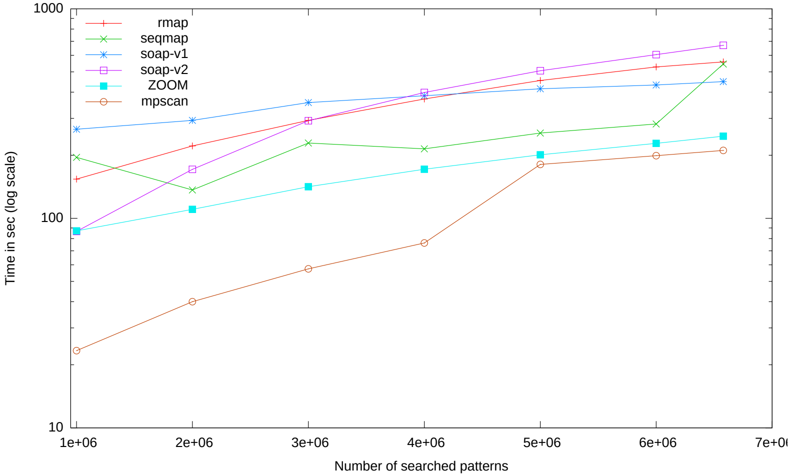
  unmapped : not found

  uniquely mapped : mapped at a single genomic location

  mutiply mapped expensive, uncomplete, complicated by repeats
- Balance: a question of tag length
  - shorter tags, more mapped tags
  - longer tags, more uniquely mapped tags

MPSCAN performance

Mapping 27 bp ChIP-seq reads on human chromosome 1



Genome annotation with SAGE/LongSAGE

1. SAGE: each 14 bp tag occurs many times in the human genome does not predict a unique location (theoretical average 12 locations)

2. LongSAGE: 21 bp high probability of a unique location [Saha et al. 02]

Genome annotation with SAGE/LongSAGE

2. LongSAGE: 21 bp high probability of a unique location [Saha et al. 02]

3. Evaluation in 2007
   on 1 million tags: 67% cannot be located
   but 80% of located tags have a unique location; [Keime et al. 07]

Genome annotation with SAGE/LongSAGE

2. LongSAGE: 21 bp high probability of a unique location [Saha et al. 02]

3. Evaluation in 2007
   on 1 million tags: 67% cannot be located
   but 80% of located tags have a unique location; [Keime et al. 07]

4. How to improve prediction of transcribed genomic regions?

Questions

- Is there an optimal tag length for prediction capacity?
- How much sequence errors with new sequencing technologies?
- How do they impact on the prediction?
- How does the prediction capacity vary with length, background distribution, and errors?
- What are the source of unmapped tags?

Methods

Mapping Background distribution

Let G be the target genome of length n, T a random Bernoulli sequence of same length. We consider tags of length t.

- Compute in function of the tag length t:
  $A(t)$: the probability of a tag not to be located in sequence T
  $B(t)$: the probability of a tag to be located once in sequence T
- Here $t \simeq \log(n)$, hence a tag should have a few locations on T.
- The law of the # of locations of a word w is approximated by a Compound Poisson distribution $\mathcal{L}_{cp}(\lambda(w), a(w))$ [Robin et al., 05]

Background Distribution mapping (II)

- The law of the # of locations of a word w is approximated by a Compound Poisson distribution $\mathcal{L}_{cp}(\lambda(w), a(w))$ [Robin et al., 05] where
  - $a(w)$ is the probability of word w to overlap itself

$$a(w) \quad = \quad \sum_{p \in Pr(w)} \mathcal{P}(w[1 \cdot \cdot p]) \quad = \quad \sum_{p \in Pr(w)} \sigma^{-p}$$

    with $Pr(w)$: *set of primary periods of* w and σ: *cardinal of the alphabet*
  - $\lambda(w)$ is the expected number of trains of w equals $(1 - a(w)) \cdot \mathbb{E}(\mathcal{N}(w))$

In the Bernoulli model:

- $\mathbb{E}(\mathcal{N}(w))$ equals n/σ^t
- $a(w)$ does not depend on w but solely on its autocorrelation c

Background Distribution mapping (III)

Average over all possible words of $a(w)$ and $\lambda(w)$

$$a = \mathbb{E}(a(c)) = \frac{\sum_{c \in \Gamma(t)} a(c) \cdot \mathcal{N}(c)}{\sigma^t} \tag{1}$$

where: $\Gamma(t)$: *set of autocorrelations of length t*
$\mathcal{N}(c)$: *population of autocorrelation c*

Computation

- Enumeration of all self-overlap vectors (autocorrelation) [Rivals & Rahmann, 03]
- Average over all classes of words with the same autocorrelation weigthed with the population of each autocorrelation

Solution

$$A(t) = e^{-\lambda} \quad \text{and} \quad B(t) = (1 - a)\lambda e^{-\lambda} \tag{2}$$

Estimation of sequence errors

- A general approach for a set of sequences, either occurrences or tags
- Biologically valid tags: those with high occurrence number
- Variables

$\mathcal{S}(t)$: the probability that a sequence of length t has at least one sequence error;

$\mathcal{X}(t)$: the prior probability that a sequence of length t is not located on G;

$\mathcal{M}(t)$: the probability that an erroneous sequence of length t is located on G;

$\mathcal{R}(t)$: the probability that a non erroneous sequence of length t is not located on G.

$$\mathcal{X}(t) = (1 - \mathcal{S}(t)) \cdot \mathcal{R}(t) + \mathcal{S}(t) \cdot (1 - \mathcal{M}(t)). \tag{3}$$

Estimation of sequence errors (II)

For a given set of experimental sequences: occurrences or tags.

$\mathcal{X}(t)$: map all sequences on G; % of seq not found

$\mathcal{R}(t)$: map biologically valid sequences on G; % of seq not found
select *valid* according to occurrence number

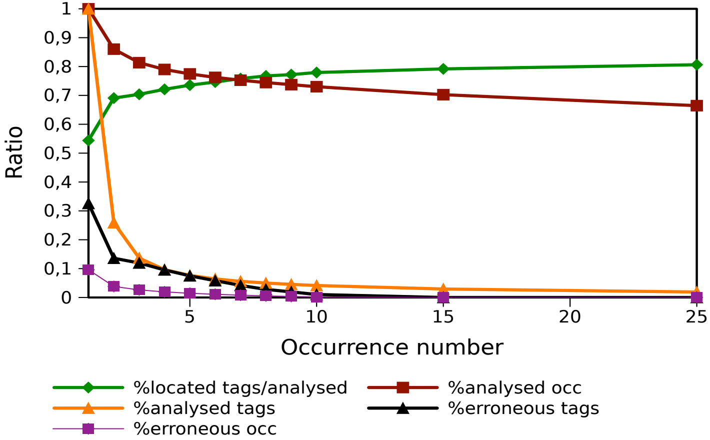
$\mathcal{M}(t)$: randomly mutate valid sequences and map them on G;
same subset as for $\mathcal{R}(t)$ % of seq found

Bootstrap: to get standard error on $\mathcal{S}(t)$

Deduce the probability of an erroneous nucleotide from that of erroneous occurrences

$$p = 1 - \exp^{\left(\frac{\log(1-\mathcal{S}(t))}{t}\right)} . \qquad (4)$$

Graphical method: choice of occurrence threshold



Data sets

a) SAGE-Sanger: collection of public LongSAGE libraries [SAGE-Genie]

$\simeq$ 9 million occurrences 1 992 500 tags at 21 bp

b) CAGE-Sanger: 5' transcriptomic tags from FANTOM3 [Kawaji et al., 06]

5 476 289 occ. for 1 627 871 tags at 21 bp

c) SAGE-Solexa private library from the Skuld-Tech® company

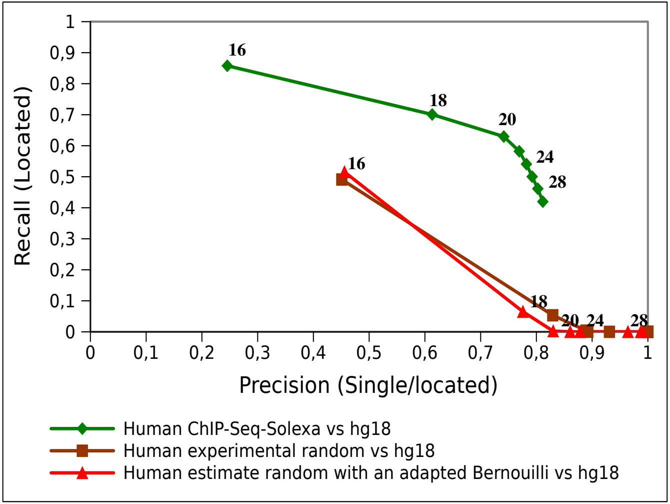
2 222 343 occurrences for 440 445 tags at 21 bp

d) ChIP-seq-Solexa from NCBI GEO sample *GSM*325935 [Barrett et al., 08]

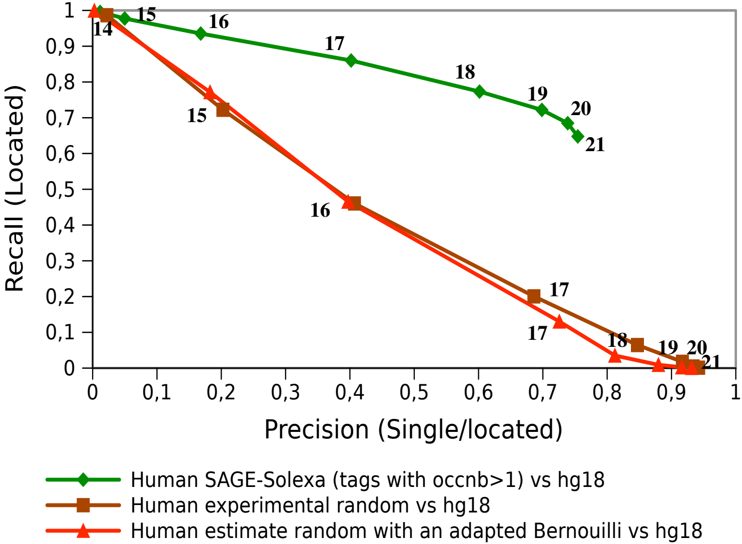
1 339 671 occ. for 929 165 tags at 30 bp

Results

Background distribution and prediction capacity for ChIP-seq



Background distribution and prediction capacity



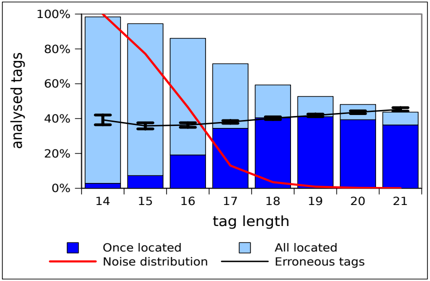
Comparative analysis of sequence errors in occurrences

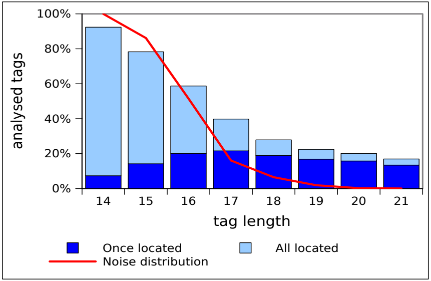
	SAGE-Sanger (6 527 650 occ)		SAGE-Solexa (2 222 344 occ)		ChIP-seq-Solexa (1 339 671 occ)	
t	$\mathcal{S}(t) \pm \alpha(t)$	p	$\mathcal{S}(t) \pm \alpha(t)$	p	$\mathcal{S}(t) \pm \alpha(t)$	p
14	6.02 ± 1.64	0.44	4.22 ± 2.77	0.31	–	–
15	6.25 ± 0.88	0.43	5.31 ± 1.26	0.36	–	–
16	6.10 ± 0.67	0.39	4.85 ± 0.96	0.31	6.89 ± 1.59	0.44
17	7.37 ± 0.46	0.45	5.24 ± 0.71	0.32	–	–
18	8.32 ± 0.38	0.48	6.65 ± 0.65	0.38	7.53 ± 0.99	**0.46**
19	**9.52 ± 0.38**	0.53	**8.11 ± 0.61**	0.44	–	–
20	10.79 ± 0.33	**0.57**	**9.14 ± 0.61**	**0.48**	**8.84 ± 0.09**	**0.48**
21	12.49 ± 0.32	0.63	10.57 ± 0.60	0.53	–	–
22	–	–	–	–	10.39 ± 0.09	0.50
24	–	–	–	–	11.99 ± 0.09	0.53
26	–	–	–	–	13.51 ± 0.09	0.56
28	–	–	–	–	15.22 ± 0.09	0.59
30	–	–	–	–	16.83 ± 0.09	**0.61**

Comparison SAGE vs CAGE

a) SAGE-Sanger (1 992 500 tags)

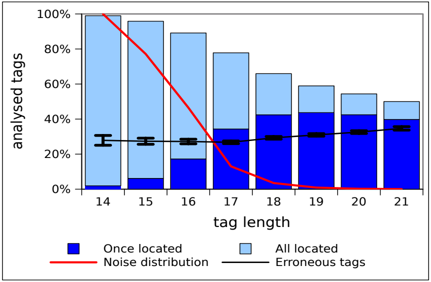
b) CAGE-Sanger (1 627 871 tags)



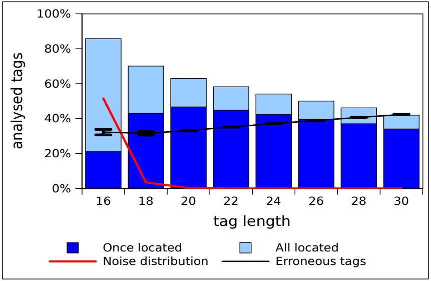


Comparison SAGE-Solexa vs ChIP-seq

c) SAGE-Solexa (440 445 tags)



d) ChIP-seq-Solexa (929 165 tags)



Annotation & comparison with tiling array

Classification of Transcriptionally Active Regions (TARs) obtained from SAGE-Solexa library according to Ensembl annotations into exonic, inxonic, intronic, and intergenic categories

Result	Total	exonic		inxonic		intronic		intergenic	
		S (1)	AS (4)	S (2)	AS (5)	S (3)	AS (6)	EST (7)	other (8)
$t = 16$	100%	34.7%	7.8%	1.0%	0.4%	15.1%	9.2%	5.5%	26.3%
	16 328	5 659	1 279	156	73	2 467	1 501	898	4 295
$t = 21$	100%	38.5%	8.8%	1.2%	0.3%	15.6%	6.6%	5.5%	23.5%
	56 006	21 600	4 947	691	192	8 760	3 694	3 054	13 068
$t = 20$	100%	38.5%	8.8%	1.2%	0.3%	15.6%	6.6%	5.5%	23.5%
	56 441	21 706	4 970	687	192	8 808	3 743	3 100	13 235
Tiling	100%	35.6%	–	–	–	34.9%	–	10.8%	18.7%

Tiling data from [Encode project, 07]

General conclusions

- Method to estimate sequence errors

  and to optimise prediction capacity in function of tag length.

- Solexa sequencing is accurate and adequate for DGE

- Probability of an erroneous nucleotide increases with its position independent of the type of assay: Digital Gene Expression or ChIP-seq

- The longer (talking about tag), may not be the better

Methodological and biological evidence

- With tags ≥ 19 bp, probability to map a position by chance $< 1\%$
- Above 20bp the number of uniquely mapped tags decreases.
- At 20bp with $\#\ occ. > 1$ the false positive rate 0.6%.
  validity of filtration
- Possibility to optimise prediction capacity with exact mapping
  by choosing a length $\simeq 20$
- SNPs affect $< 4.6\%$ of the tags
- 9.6% of transcriptomic tags are not mapped due to artefactual or biological reasons

Future work

- Bioinformatic platform for the analysis of transcriptomics & epigenomics assays: routine analysis
- Database of transcriptomic tags and annotations
  for each tag: genomic location and related annotations
- Background distribution for a markov model of the genome sequence
- Approximate mapping with a few mismatches
- Extension for longer reads and other applications:
  genotyping, breakpoint mapping [Chen et al., 08],
  genome resequencing [Dohm et al., 08], metagenomics

Authors and acknowledgments

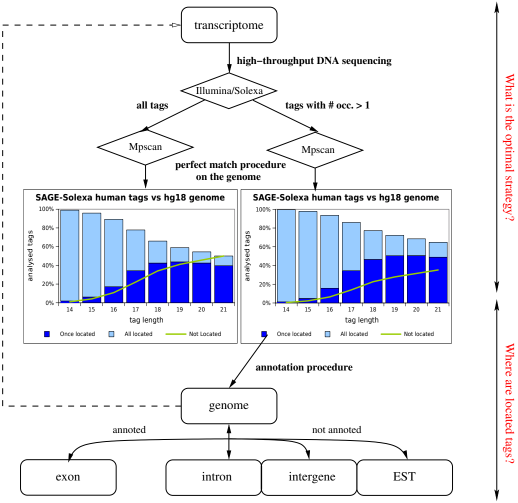
- L.I.R.M.M., Montpellier
  N. Philippe, L. Bréhélin, E. Rivals
- Helsinki University of Technology, Finland
  Jorma Tarhio
- Institut de Génétique Humaine (I.G.H.), Montpellier
  A. Boureux, Thérèse Commes, Groupe Etude des Transcriptomes

Thanks to:

- Skuld-Tech® Montpellier
  D. Piquemal for SAGE-Solexa library and data
- S. Schbath, MIG INRA Jouy-en-Josas
- BioMIPS Languedoc Roussillon, Ligue Régionale contre le Cancer
- Cancéropôle Grand Sud Ouest

Thanks for your attention

Strategy schema



transcriptome

high−throughput DNA sequencing

Illumina/Solexa

all tags

tags with # occ. > 1

Mpscan

Mpscan

perfect match procedure
on the genome

annotation procedure

genome

annoted

not annoted

exon

intron

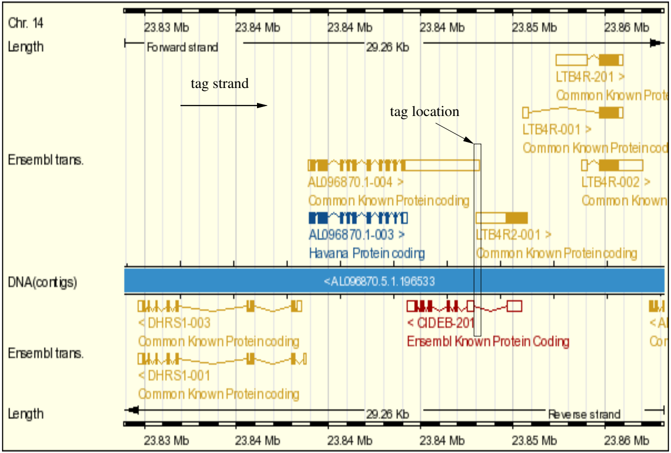
intergene

EST

What is the optimal strategy?

Where are located tags?

Tag annotation is difficult


Chr. 14
Length
Ensembl trans.
DNA(contigs)
Ensembl trans.
Length
23.83 Mb
23.84 Mb
23.84 Mb
23.84 Mb
23.85 Mb
23.86 Mb
Forward strand
29.26 Kb
tag strand
tag location
LTB4R-201 >
Common Known Prote
LTB4R-001 >
Common Known Protein cod
AL096870.1-004 >
Common Known Protein coding
LTB4R-002 >
Common Known
AL096870.1-003 >
Havana Protein coding
LTB4R2-001 >
Common Known Protein coding
< AL096870.5.1.196533
< DHRS1-003
Common Known Protein coding
< CIDEB-201
Ensembl Known Protein Coding
< Al
Cor
< DHRS1-001
Common Known Protein coding
29.26 Kb
Reverse strand
23.83 Mb
23.84 Mb
23.84 Mb
23.84 Mb
23.85 Mb
23.86 Mb