Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity

Nicolas Philippe, Anthony Boureux, Laurent Brehelin, Jorma Tarhio, Thérèse Commes, Eric Rivals

Published online 16 June 2009

Nucleic Acids Research, 2009, Vol. 37, No. 15 ***e104***
doi:10.1093/nar/gkp492

Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity

Nicolas Philippe[1,2], Anthony Boureux[2], Laurent Bréhélin[1], Jorma Tarhio[3], Thérèse Commes[2] and Éric Rivals[1,*]

[1]Laboratoire d'Informatique, de Robotique et de Microélectronique, Université de Montpellier II, UMR 5506 CNRS, 161 rue Ada, 34392 Montpellier, France, [2]Groupe d'études des transcriptomes, Université de Montpellier II, Institut de Génétique Humaine, UPR1142-CNRS, Place Eugène Bataillon, 34095 Montpellier, France and [3]Helsinki University of Technology, PO Box 5400, FI-02015 HUT, Finland

Received January 30, 2009; Revised May 19, 2009; Accepted May 21, 2009

ABSTRACT

Ultra high-throughput sequencing is used to analyse the transcriptome or interactome at unprecedented depth on a genome-wide scale. These techniques yield short sequence reads that are then mapped on a genome sequence to predict putatively transcribed or protein-interacting regions. We argue that factors such as background distribution, sequence errors, and read length impact on the prediction capacity of sequence census experiments. Here we suggest a computational approach to measure these factors and analyse their influence on both transcriptomic and epigenomic assays. This investigation provides new clues on both methodological and biological issues. For instance, by analysing chromatin immunoprecipitation read sets, we estimate that 4.6% of reads are affected by SNPs. We show that, although the nucleotide error probability is low, it significantly increases with the position in the sequence. Choosing a read length above 19 bp practically eliminates the risk of finding irrelevant positions, while above 20 bp the number of uniquely mapped reads decreases. With our procedure, we obtain 0.6% false positives among genomic locations. Hence, even rare signatures should identify biologically relevant regions, if they are mapped on the genome. This indicates that digital transcriptomics may help to characterize the wealth of yet undiscovered, low-abundance transcripts.

INTRODUCTION

Unravelling the complexity of the human transcriptome remains a major challenge, especially given the importance of transcriptional activity in non-coding regions of the genome (1). By applying high-throughput sequencing technologies (HTS) to open Digital Gene Expression (DGE), it is possible to catalogue all transcripts and measure their activity level in a cell (2), while finding their region of origin on the genome opens the way to their complete functional annotation. The latter is achieved if, when mapping the sequence signature (also called tag) of a transcript, one points to a single genomic location (3–5). Indeed, multiple locations are more complex to annotate and usually discarded. For a given amount of sequencing, using short tags allows a deeper sampling and a larger catalogue, while longer signatures increase the probability of detecting a single location of origin for each transcript. This has been previously suggested in a seminal work, which proposed to use 21 bp instead of the classical 14 bp tags with Serial analysis of Gene Expression method (SAGE), to annotate genomic transcribed regions (6). Today, HTS can produce longer tags (up to 36) at a much larger scale. Hence, the question of an optimal tag length is timely. However, recent investigations have provided evidence that a large proportion of tags cannot be mapped on the genome (3,7).

Assays are now designed to explore the transcriptome or more specifically the RNA dark matter, as deeply as possible and generate huge tag sets (2,8). Presently, the *prediction capacity*, i.e. the ability to identify single genomic locations for a maximum number of transcripts, has never been evaluated. Several parameters such as the tag

*To whom correspondence should be addressed. Tel: +33 4 67 41 86 64; Fax: +33 4 67 41 85 00; Email: rivals@lirmm.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

length, the background distribution (i.e. the probability of mapping a random sequence) and sequence errors influence the outcome; these factors should therefore be assessed and taken into account in the strategy.

Chromatin immunoprecipitation by sequencing (ChIP-Seq) is a method for identifying the chromosomic locations where a given protein binds to DNA. The DNA fragments obtained by immunoprecipitation are then sequenced with HTS (9,10). As for DGE, reads generally range between 25 and 34 bp and are mapped back to the genome sequence to get a genomic profile. Here again, we face the same issues as with transcriptomic assays. Moreover, sequence-based assays represent, in both cases, appealing alternatives to those based on hybridization: they overcome the difficulties inherent to this type of platforms and to hybridization, but they also offer important advantages like speed, small extract needs, and low cost (2,4). Undoubtedly, these techniques will deeply impact biological investigations and it is crucial to thoroughly evaluate their prediction capacities.

We thus propose a computational protocol to investigate how the prediction capacity of a sequence-based assay is influenced by several factors: read or tag length, background distribution, and sequence errors (hereafter, we only use the word tag). Our protocol based on exact mapping of increasing sub-parts of the tags on a reference genome sequence is illustrated in Supplementary Figure 1. It consists of a theoretical model of background distribution as a function of the tag length, a procedure to estimate sequence errors and allows one to identify which tag length optimizes the prediction capacity. Effective predictions are then validated by intersecting putative transcribed regions with Ensembl annotations, and comparing these with those yielded by a whole genome tiling array (11). We apply our protocol to both DGE and ChIP-Seq datasets, as well as to data produced with Sanger and Solexa sequencing techniques. Our analysis gives several clues to improve predictions obtained with such experiments. Finally, it provides the first independent estimates of the sequence errors per tag and per position, which indicates how error probability varies with tag length, and it measures the impact of Single Nucleotide Polymorphism (SNPs) and other biological processes on tag sets.

MATERIALS AND METHODS

Datasets

In experiments like SAGE, ChIP-Seq, etc., one obtains a collection of sequences of length t, with some being observed several times. For clarity, we call a representative sequence a *tag*, and all sequences identical to that tag observed in the collection are termed its *occurrences* (*occ.*). The *number of occurrences* (# *occ.*) of a tag, i.e. the number of times its sequence is observed in the collection, is considered as a measure of its biological validity: a tag observed once (# *occ.* = 1) may be an artefact, while a tag observed say 10 times (# *occ.* = 10) is likely a valid biological observation (2).

The data were collected from publicly available repositories (June 2008): for the human SAGE-Sanger dataset, we used the libraries from GPL1485 and GPL5624 Gene Expression Omminbus (GEO) Platforms (http://www.ncbi.nlm.nih.gov/geo) and all Sanger sequenced libraries from the CAGP project (Sage genie ftp://ftp1.nci.nih.gov/pub/SAGE/). For the human SAGE-Solexa dataset, we used a private library sequenced with the Solexa sequencer from the Skuld-Tech® company (which contains 2 222 343 occurrences for 440 445 tags). For the ChIP-Seq dataset, we used data from the GSM325935 GEO sample (rep3 - lane B), which contains 1 339 671 occ. for 929 165 tags. For the CAGE dataset, we used data from Kawaji *et al.* (12) (5 476 289 occ. for 1 627 871 tags at 21 bp only). The human genome (hg18, NCBI Build 36.1) was retrieved from the UCSC Genome Browser website (http://genome.ucsc.edu).

Locating tags

To predict regions of origin of experimental tags in the human genome, we searched for these tags in the genome sequence with MPSCAN (13), and recorded whether each tag has been located or not. MPSCAN is guaranteed to report, for each tag, all positions at which the genome exactly matches the tag. Such positions are called *genomic locations* or simply *locations*. We distinguish tags found once in the genome, i.e. *uniquely mapped* tags, from those that *map multiple locations*, i.e. *multi-mapped* tags. To investigate how the proportion of uniquely mapped tags varies according to the tag length, we searched for prefixes or suffixes of increasing lengths of experimental tags, e.g. for a 21-bp tag, we also searched for its prefixes of length 14, 15, ... until 20.

To illustrate the tradeoff between uniquely and multi-mapped tags as a function of length, we plot a *Precision–Recall-like* curve in which the *Recall* is the percentage of mapped tags over all tags, while the *Precision* is the percentage of uniquely mapped tags over all mapped tags.

Background distribution

We introduce here the notation used throughout the article. Consider a target genome G of length n (by default, the human genome in this work). To optimize the annotation strategy with respect to tag length we need to compute the probability that a random tag w of length t has 0 or 1 matching locations on a random sequence T of length n. We consider random sequences under the Bernoulli model. In our context, tags are short compared with the genome, but long enough to match only a few locations on the genome, since t is on the order of $\log(n)$. We describe the theoretical probabilities for unconstrained tags (general model), as well as for tags starting with a defined prefix (adapted model; i.e. CATG for SAGE like tags).

General model. Precisely, we need to determine the following two probabilities: 1/ $A(t)$: the probability that a tag w of length t is not located on sequence T, and 2/ $B(t)$: the probability that a tag w of length t is located exactly *once* in T. If $\mathcal{N}_T(w(t))$ denotes the number of matching locations of w in T, we have

$$A(t) := \mathbb{P}(\mathcal{N}_T(w(t)) = 0) \text{ and } B(t) := \mathbb{P}(\mathcal{N}_T(w(t)) = 1) \quad \mathbf{1}$$

It is known that the probability distribution of $\mathcal{N}_T(w(t))$ is precisely approximated by a compound Poisson distribution $\mathcal{L}_{cp}(\lambda, a)$, where a is the probability of a tag location overlapping the previous location, and λ is the expectation of the number of consecutive overlapping locations (14). Hence, one obtains

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

The difficulty is to compute over all possible tag sequences an average of a and of λ, while accounting for the self-overlap possibilities of all tags, i.e. their period sets. For this sake, we use an algorithm to enumerate all period sets efficiently (15).

Adapted model. This model can be adapted to the case of tags starting with a predefined prefix. We consider the case of NIaIII SAGE tags, which start with a CATG restriction site. Hence, the number of matching positions is limited to locations where CATG appears in the target genome. Let n_{site} denote their number. To estimate our probabilities, that we denote by $A'(t)$ and $B'(t)$ in the adapted model, we simply consider that the random sequence T contains as many CATG sites as the target genome (i.e. n_{site}). This number replaces n in the formula of parameter λ, then $A'(t)$ and $B'(t)$ are computed as in Equation (2). To verify that these formulas are precise enough for our purpose, we also empirically estimated $A'(t)$ and $B'(t)$ by mapping randomly generated tags to a random sequence. The comparison between theoretical and empirical values of $A'(t)$ and $B'(t)$ is shown in Figure 1B.

In the following, $1-\mathcal{A}(t)$ (resp. $1-\mathcal{A}'(t)$) represents the *background probability of being mapped* in the general model (resp. in the adapted model).

Estimating sequence errors

Given a collection of sequences (either the tags or occurrences), it is possible to estimate the number of sequence errors of an experiment as a function of the sequence length t. We present a framework and an algorithm to estimate the probability that a sequence of length t is erroneous. If the sequence is an occurrence, it is erroneous if it contains a sequence error, while a tag is erroneous if all of its occurrences are erroneous. Thus, a non-erroneous tag has at least one non-erroneous occurrence. Later, we apply this algorithm both to the set of tags and to the collection of occurrences. Let us consider the following probabilities:

- $\mathcal{S}(t)$: 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.

We want to compute $\mathcal{S}(t)$. Now, why can a sequence not be found on the genome?

(i) Either, it is a biologically valid sequence that is not erroneous, and does not appear in the genome for biological reasons (i.e. post-transcriptional modification, SNP); the probability of this event is given by $(1-\mathcal{S}(t)) \cdot \mathcal{R}(t)$;
(ii) or it is an erroneous sequence, which does not appear in the genome, and in this case the probability is $\mathcal{S}(t) \cdot (1-\mathcal{M}(t))$.

As both cases are exclusive, the probability $\mathcal{X}(t)$ that a sequence is not located on G is

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

Now, if we can estimate $\mathcal{M}(t)$, $\mathcal{R}(t)$ and $\mathcal{X}(t)$, we will be able to estimate $\mathcal{S}(t)$ using the formula:

$$\mathcal{S}(t) = \frac{\mathcal{X}(t) - \mathcal{R}(t)}{1 - \mathcal{M}(t) - \mathcal{R}(t)}. \quad \mathbf{4}$$

We explain how to estimate these probabilities. Clearly, $\mathcal{X}(t)$ can be estimated from the number of experimental sequences that are not located on G. For $\mathcal{R}(t)$, we need a sample of non-erroneous sequences. Assuming that above a certain threshold of # *occ.* a tag is biologically valid, we choose a threshold by a graphical method presented in 'Sequence errors' Section (see the blue curve in Figure 2) and select tags whose # *occ.* lies above that threshold. We estimate $\mathcal{R}(t)$ as the proportion of these sequences that are not located on the genome. For $\mathcal{M}(t)$, we assume that any erroneous sequence contains no more than one error (a likely hypothesis as the nucleotide error probability is expected to be low) and proceed by simulation: we consider the same collection of sequences as for $\mathcal{R}(t)$, randomly substitute one position in each sequence and search it on G. The proportion of sequences not found in G is our estimate for $\mathcal{M}(t)$.

We evaluate the precision of our estimate $\mathcal{S}(t)$ as a function of t by computing a standard error $\alpha(t)$ for $\mathcal{S}(t)$ using the bootstrap technique (16). For each of the k bootstrap samples ($i := 1..k$), we recompute all the statistics and $\mathcal{S}_i(t)$ denotes the $\mathcal{S}(t)$ obtained with bootstrap sample i. In our experiments, we use $k = 100$. The standard error is then computed with

$$\alpha(t) = \sqrt{\frac{\sum_{i=1}^{k} (\mathcal{S}_i(t) - \mathbb{E}[\mathcal{S}_i(t)])^2}{k-1}}. \quad \mathbf{5}$$

From the $\mathcal{X}(t)$, $\mathcal{M}(t)$ and $\mathcal{S}(t)$ estimated on the tags, one can also compute the proportion of located tags that are erroneous using the following formula:

$$\mathcal{V}(t) = \frac{\mathcal{S}(t) \cdot \mathcal{M}(t)}{1 - \mathcal{X}(t)}. \quad \mathbf{6}$$

Note that $\mathcal{V}(t)$ should not be confused with $\mathcal{M}(t)$, the probability that an erroneous sequence is located.

Moreover, from the $\mathcal{S}(t)$ estimated on the occurrences, we can estimate the sequence error at the nucleotide level. Let p be the error probability for one nucleotide. The probability that an occurrence of length t has no error is $(1-p)^t$. Thus, $\mathcal{S}(t) = 1-(1-p)^t$,

and we get

$$p = 1 - \exp\left(\frac{\log(1 - \mathcal{S}(t))}{t}\right). \quad 7$$

Classification of transcriptomic tags

For transcriptomic tags, we consider the subset of uniquely mapped tags, and given their genomic location, determine if they fall in a region annotated by a gene or an EST according to Ensembl. This can be a complex matter if genes are nested within each in another, or on both strands. We classify the tags giving a higher priority to gene versus EST annotations, and to annotations on the same, rather than on the opposite, strand. The classification algorithm proceeds as follows. Relative to its strand, if the tag is located in a gene, the tag is *exonic* (1) if it falls entirely into an exon, or *inxonic* (2) if it covers an intron–exon border or *intronic* (3). Otherwise, the same is done with a gene on the opposite strand (if any), which yields the cases exonic (4), inxonic (5) and intronic (6). Then, if it is still not annotated, we check whether it is covered by an *EST* on any strand (7), and otherwise the tag is said to be *intergenic* (8).

RESULTS

Background distribution, recall/precision tradeoff and optimal tag length

To determine the appropriate tag length for genomic annotation, it is necessary to measure the background distribution when mapping tags to the genome. For this, we computed as a function of tag length the probabilities that a tag 1/ is mapped on the genome at least once $[1 - A'(t)$, with t denoting the length], 2/ matches a single location $[B'(t)]$, under a Bernoulli model. We approximate the law of these probabilities using the guaranteed Poisson approximation, which provides more precise estimates than published so far (6). For the human genome case, we report these probabilities for both transcriptomic tags starting with a pre-defined prefix (CATG for SAGE-like tags) in Figure 1A and for unconstrained tags (ChIP-Seq, RNA-Seq) in Supplementary Figure 2A. For instance, at length 16 the probability that a transcriptomic tag is found at least once in a random sequence is 46%. This probability measures a background distribution: 46% of tags that are located on a real genome would also be located on a random sequence of the same length. This can be understood as a *P*-value (as in a BLAST result): the lower the probability, the less likely to find the same result by chance. Clearly, one gains confidence in the result of a mapping procedure by choosing a tag length for which this probability is low. From Figure 1A, the probability of being mapped in a random sequence logically decreases from nearly 1 at 14 bp to 10^{-6} at 25 bp. To keep the background probability of being mapped <1%, one should use tags longer than 18 bp. The probability of being located once, $B'(t)$, (i.e. to map a single location) increases up to 16 bp, then decreases until 25 bp and converges towards zero as $[1 - A'(t)]$.

Of course, the tag length should be chosen to minimize the probability of being unmapped and maximize that of being mapped once. The balance between these antagonistic goals can be assessed on a *Precision–Recall-like* curve, which plots (for all lengths) the *Recall*, i.e. the percentage of mapped tags over all tags, versus the *Precision*, i.e. the percentage of uniquely mapped tags over mapped tags (Figure 1B). In addition, to check whether our probabilistic model is adequate, we estimated the same probabilities by mapping randomly generated tags. The *Precision–Recall-like* curves for theoretical and empirical random mapping are compared with that of experimental transcriptomic tags in Figure 1B. First, the curves of theoretical and empirical estimations (orange and brown) are closely correlated, suggesting that both the model and its adaptation to constrained transcriptomic tags are correct. Second, the curve of true tags (blue) departs from the random expectation and illustrates the tradeoff between recall and precision: it is linear from 14 until 19 bp and then bends down. Beyond 19 bp, increasing the tag length induces a loss in recall that is not compensated by a gain in precision. To adjust the length according to both criteria,

A

t \ $\mathbb{P}$	$1 - A'(t)$	$B'(t)$
	Located	Once located
14	0.9997	0.0017
15	0.7717	0.1409
16	0.4654	0.1849
17	0.1303	0.0946
18	0.0351	0.0285
19	0.0091	0.0080
20	0.0023	0.0021
21	$5.93\ 10^{-5}$	$5.53\ 10^{-5}$
25	$2.44\ 10^{-6}$	$2.3\ 10^{-6}$

B

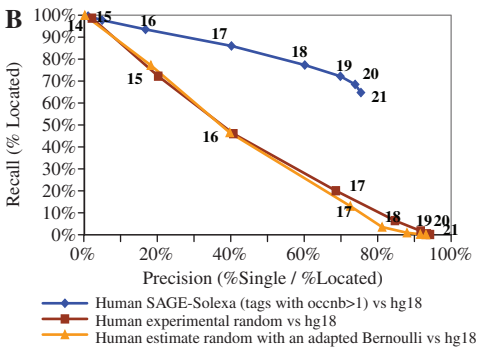

Figure 1. Background distribution and influence of length on the prediction capacity. (**A**) Theoretical probabilities of a transcriptomic SAGE tag being located $(1 - A'(t))$, and being located once $(B'(t))$ in a random Bernoulli sequence. The former starts at approximately 1 at 14 bp and gets to lesser than 0.01 at 19 bp, while the latter reaches its maximum at 18 bp and decreases towards low values. (**B**) Influence of the tag length onthe prediction capacity shown with *Precision–Recall-like* curves. The recall (percentage of located tags over all tags) is plotted versus the precision (percentage of uniquely mapped tags over all mapped tags) for each tag length. The blue curve of a human Solexa tags set departs from those of random tags, either theoretically computed (yellow) or empirically estimated (brown curve). The overlapping yellow and brown curves show the validity of thetheoretical model. The blue curve remains linear until 19 bp, then bends down.

the background probability and the tradeoff between precision and recall, one could choose an optimal length at 19–20 bp, for which both the recall and precision are above 70%, and the background probability is <1%. We obtain similar results for unconstrained tags, as shown in Supplementary Figure 2B. The method can be used for any genome and any type of tags. The *Precision–Recall-like* curve allows one to adjust the tag length to its experimental set up.

Sequence errors

From the results of mapping a set of experimental tags on a genome, one can estimate the number of sequence errors produced in the assay. We propose a procedure to compute the probability that one occurrence or one tag contains at least one error in its sequence. For this we empirically measure the probability that a biologically valid sequence is not mapped, and the probability of an erroneous tag is mapped. We apply our method to estimate the sequence errors in transcriptomic and ChIP-Seq assays obtained with Solexa and Sanger sequencing techniques using real datasets. We report the percentages of both erroneous occurrences (Table 1) and tags (Supplementary Table 3), and derive from the former the percentage of erroneous nucleotides (Table 1).

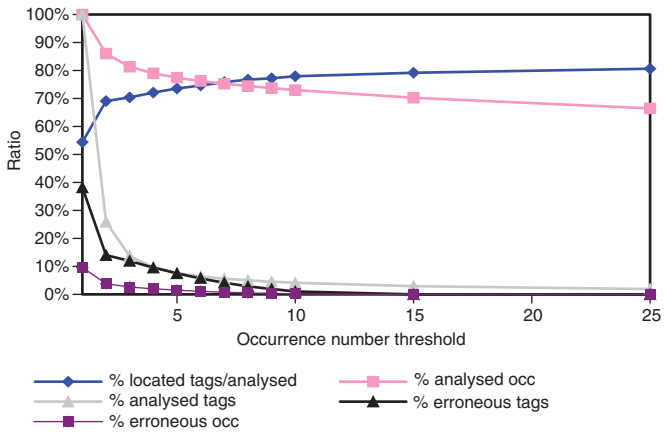

Figure 2. Influence of the selection of tags with # *occ.* above a threshold on the percentage of analysed tags (grey), analysed occurrences (pink), erroneous tags (black), erroneous occurences (magenta) and located tags (blue). A point at abcissa x is the corresponding value when only tags with # *occ.* $> x$ are kept. Analysed tags represent 25% of the original tag set, but still 85% of the occurrences. The percentages of erroneous tags and erroneous occurrences become low, respectively very low with # *occ.* > 1. The percentage of located tags stabilizes after # *occ.* > 10; the blue curve serves as a graphical method to set the threshold to select biologically valid tags in our estimation procedure.

Note that for estimating $\mathcal{R}(t)$ and $\mathcal{M}(t)$, as biologically valid tags we select those whose occurrence number lies above an empirical threshold of 10. We choose the threshold by plotting the proportion of located tags as a function of the # *occ.* (blue curve in Figure 2). As the probability of an erroneous tag decreases with the # *occ.*, the curve should level off above a given threshold. For a SAGE-Solexa assay, the blue curve shows that the proportion of mapped tags is almost stable above # *occ.* $\geq$10.

We also investigate the effect of solely considering the tags whose # *occ.* lies above a user-defined threshold on the proportions of analysed and erroneous tags and occurrences. For each threshold value, we recomputed all of this information and plotted it on Figure 2. Note that both the levels of erroneous occurrences and tags drop drastically (divided by 3) when considering only # *occ.* > 1. Thus, the usage of selecting tags whose # *occ.* > 1 for further analysis ('Annotation' section) is appropriate since although they account for $\simeq$25% of tags (grey curve), they represent >85% of the occurrences (pink curve). This filtration also positively affects the probability that a tag with a mapped location is due to an erroneous tag, i.e. the rate of *false positive locations*. Indeed, for the SAGE-Solexa tags, $\mathcal{V}(20)$ drops from 2.17 to 0.58% with the filtration (Table 2).

Estimates of the proportion of erroneous occurrences are less precise below 16 bp as shown by the

Table 1. Percentages of erroneous occurrences [$\mathcal{S}(t)$] and the probability of an erroneous nucleotide (P) for SAGE and Chip-seq assays at different tag length (t)

	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

The tag length t ranges from 14 to 21 for SAGE-{Sanger,Solexa}, and from 16 to 30 bp for Chip-Seq-Solexa, $\pm\ \alpha(t)$ is the standard error of $\mathcal{S}(t)$. The percentage of erroneous occurrences logically increases with length. However, the percentage of an erroneous nucleotide increases with its position in the sequence (until the 30th bp), showing that, even with the Solexa technique, errors occur more frequently at the 3′ end (in bold: values cited in text).

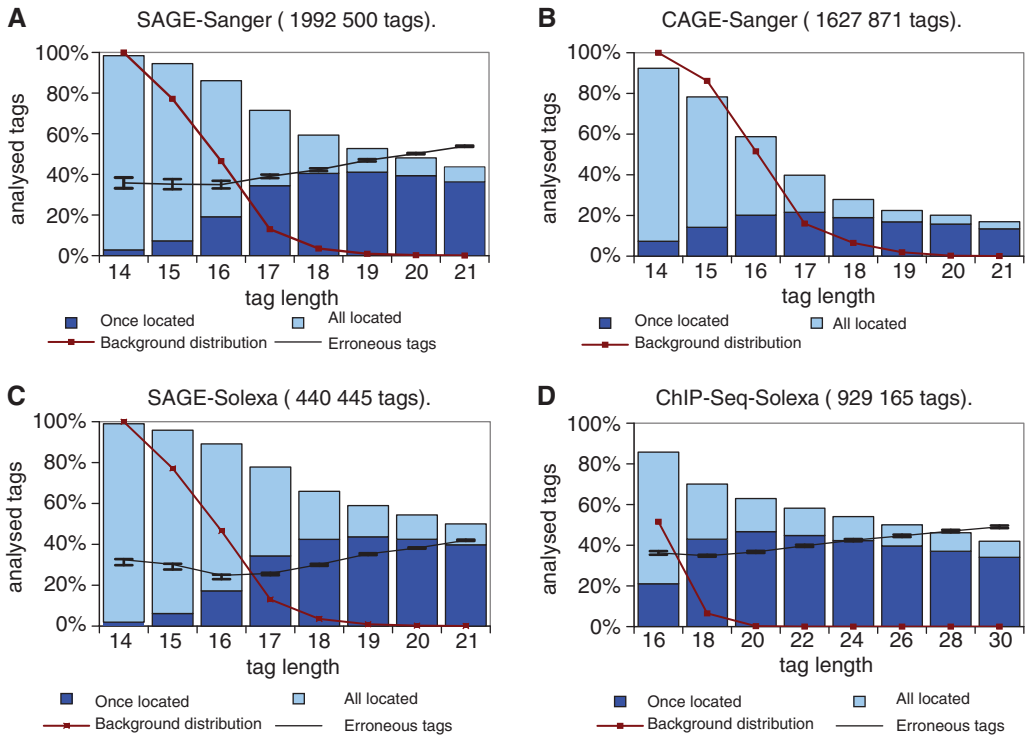

Figure 3. Variation of the prediction capacity when mapping four transcriptomic or ChIP-Seq tag setson the human genome. For each assay, the histogram gives for each tag length, the percentages over all tags of located tags (light blue bar) uniquely mapped tags (dark blue bar), erroneous tags with the standard errors (black curve) and the background distribution (brown curve). Histograms obtained whenmapping (**A**) the SAGE-Sanger set, (**B**) the CAGE-Sanger set, (**C**) the SAGE-Solexa set and (**D**) the ChIPSeq-Solexa set on the genome sequence (hg18). For concision, only even lengths are plotted in ChIP-Seq histogram (D) tags. The ratio of tags located once on the genome reaches its maximum at 19 bp when the background probability of being located is already low, while the ratio of erroneous tags keeps on increasing after that.

standard errors. This is due to the background probability of mapping tags of these lengths in a random sequence. For all types of assays, the percentage of erroneous occurrences increases regularly above a certain length, which is expected since the probability of having at least one error increases with the tag length. For the SAGE-Solexa assay, the percentage decreases until 16 bp and then increases regularly up to 21 bp. For the ChIP-Seq-Solexa assay, it also increases until 30 bp, thus confirming the trend noted with SAGE, with similar error levels (8.84 versus 9.14 for SAGE at 20 bp). Hence, tags above 20 bp may not be the most appropriate for annotating the genome.

Above 16 bp, SAGE-Solexa exhibits a lower error rate than with the Sanger sequencing method: 8.11% of occurrences are erroneous at 19 bp with Solexa, while it reaches 9.52% with the Sanger technique. This confirms that high-throughput sequencing technologies are well adapted for digital transcriptomic assays.

The percentage p of erroneous nucleotides is derived from that of erroneous occurrences (but not of erroneous tags, since the occurrences are sequenced). The results for all three assay types are given in Table 1. The error level obtained with the Sanger method agrees with published estimates (17,18). Note that the nucleotide error rate increases with tag length for all assays. For ChIP-Seq-Solexa, the error rate goes from 0.46% at 18 bp to 0.61% at 30 bp: an increase of 32%. Hence, the probability of getting an error on a nucleotide increases with the nucleotide position. Again, this suggests that, when tags are used for genome annotation, the longest is not the best.

Comparison of the mapping capacity of assays and sequencing techniques

Having estimated the background distribution and the error levels, we can draw our attention back to our main question: how well can one predict genomic sites of interest using assays based on short sequences? In other words, what is the capacity to map single genomic sites with experimental tags and how do several factors (background distribution, error, tag length or sequencing technology) affect this capacity? Here, we investigate this issue for two types of transcriptomic (CAGE and SAGE) and one of epigenomic (ChIP-Seq) assays, as well as for two sequencing techniques (Solexa versus Sanger). To each case corresponds in Figure 3 a histogram, which gathers all the information for various tag lengthes. This way, by comparing two adjacent sub-figures, one can evaluate the effect of either the assay (e.g. CAGE 3.B versus SAGE 3.A), or the sequencing technique (Solexa 3.C versus Sanger 3.A).

Let us first describe features common to all cases. First, the percentage of mapped tags (light blue) necessarily decreases with the tag length t since shorter tags are

Table 2. Errors and localization for the Sage-Solexa library without and with filtration

Sage-Solexa private library			# *occ.* > 0 (440 445 tags)			# *occ.* > 1 (114 721 tags)		
t	$\mathcal{R}(t)$	$\mathcal{M}(t)$	$\mathcal{X}(t)$	$\mathcal{S}(t) \pm \alpha(t)$	$\mathcal{V}(t)$	$\mathcal{X}(t)$	$\mathcal{S}(t) \pm \alpha(t)$	$\mathcal{V}(t)$
14	0.23	97.36	1.01	32.46 ± 2.80	31.92	0.59	14.77 ± 2.82	14.47
15	0.96	88.42	4.16	30.18 ± 1.74	27.84	2.25	12.22 ± 1.75	11.06
16	2.34	63.35	10.86	24.84 ± 1.31	17.65	5.86	10.27 ± 1.32	6.91
17	6.84	33.65	22.18	25.77 ± 0.77	11.14	13.50	11.19 ± 0.78	4.35
18	10.27	10.83	34.06	30.15 ± 0.77	4.95	20.09	12.45 ± 0.77	1.69
19	12.38	6.49	41.04	35.33 ± 0.74	3.89	23.61	13.85 ± 0.74	1.18
20	13.89	3.09	45.61	38.20 ± 0.79	**2.17**	25.52	14.01 ± 0.79	**0.58**
21	15.56	2.42	50.00	41.98 ± 0.99	2.04	28.23	15.48 ± 0.99	0.57

For each length t, one reads the percentages over all tags of valid mapped tags [$\mathcal{R}(t)$], erroneous mapped tags [$\mathcal{M}(t)$], unmapped tags [$\mathcal{X}(t)$], erroneous tags [$\mathcal{S}(t)$] with its standard error [$\alpha(t)$] and false positive locations [i.e locations mapped by erroneous tags, $\mathcal{V}(t)$]. The three last statistics are given for the unfiltered (# *occ.* > 0) and filtered (# *occ.* > 1) tag set (in bold: values cited in text).

Table 3. Classification of TARs according to Ensembl annotations

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
	N	16 328	5659	1279	156	73	2467	1501	898	4295
$t = 21$	%	100	38.5	8.8	1.2	0.3	15.6	6.6	5.5	23.5
	N	56 006	21 600	4947	691	192	8760	3694	3054	13 068
$t = 20$	%	100	38.5	8.8	1.2	0.3	15.6	6.6	5.5	23.5
	N	56 441	21 706	4970	687	192	8808	3743	3100	13 235
Tiling	%	100	35.6	–	–	–	34.9	–	10.8	18.7

The number and percentage of all TARs using digital transcriptomic tags at length 16, 20 or 21 bp, and using a tiling-array [ENCODE project (11)] are shown for each annotation category (cf. 'Classification of transcriptomic tags' section). If inside a gene, a TAR can be located in an exon, in an intron or overlap one of each, we term it as 'inxonic'. These three categories are further subdivided into sense (S) and antisense (AS), depending on which strand the tag was located on compared with the gene. Category (7) concerns ESTs that do not overlap any annotated exon.

prefix of longer ones in our protocol. Second, as uniquely mapped tags are a subset of all mapped tags, the dark blue bar can at most fill the light blue bar. It first increases with t, reaches a maximum around 19 or 20 bp (17 for CAGE), and then decreases with t. The background probability of being mapped decreases sharply from nearly 100% at 14 bp to ≃15% at 17 bp, then more smoothly goes <1% at 19 bp; this occurs regardless of the type of tags (unconstrained or constrained), although the random model differs ('Background distribution, recall/precision tradeoff and optimal tag length' section). The error level (which was not evaluated for CAGE) measured as the percentage of erroneous tags increases gradually after a certain length (16 or 17 bp) at which the background probability of being mapped becomes low. Note that even if percentage of erroneous tags (and not of occurrences) seems high, the errors are made on occurrences and represent a few base pairs per Kilo base pairs. This is why most of erroneous tags have an # *occ.* = 1 and can be readily excluded from the analysis (cf. Table 2). Globally, all sub-figures except for CAGE share the same behaviour regarding the mapping capacity (note that for ChIP-Seq only even tag lengths are plotted).

SAGE-Sanger versus CAGE-Sanger. Clearly, in Figure 3 the CAGE assay departs from all other cases including SAGE-Sanger: the percentage of mapped tags at 19 bp is less than half that of LongSAGE. This strange behaviour is due to the presence of CAGE-specific errors, so our protocol is unsuitabled in this case. Indeed, 88% of all CAGE tags start with a guanine, a percentage too biased not to be artefactual. We adapted our protocol by mapping increasing tag suffixes instead of prefixes. Then, at length 18, one maps 46% of the tags versus 28% with prefixes. This confirms that errors are generated at the 5′-end of CAGE tags (19), and that one should consider a specific trimming procedure before further analysis.

SAGE-Solexa versus SAGE-Sanger. Here, we compare the results obtained with a single human LongSAGE library sequenced with the Solexa technique (≃440 K tags, 2.2 M occ.) to those yielded by the collection of all LongSAGE libraries sequenced with the Sanger method (1.9 M tags, 6.5 M occ.). In both cases, curves representing background distributions are identical. As the LongSAGE protocol has been simplified with new sequencing techniques, we observed the cumulated effect of the differences in assays and sequencing techniques. However, the two behaviours are quite similar (3.A versus 3.C), with SAGE-Solexa having a slight advantage in terms of mapping capacity: the percentage of mapped tags (light blue)

ranges from 66 to 50% for lengths 18 to 21 bp for Solexa, and from 59 to 43% for Sanger over the same length range. The difference in overall capacity, e.g. 59% versus 48% at length 19, corresponds exactly to the difference in percentage of erroneous tags 35% versus 47% for SAGE-Sanger (cf. Supplementary Table 3). This may well be due to differences in sampling depth and cumulative analysis of multiple SAGE-Sanger libraries.

In both studies, the maximum of uniquely mapped tags (dark blue) is reached at length 19: 44% for Solexa versus 41% for Sanger, again suggesting that 19 bp is a more advantageous length for genome annotation.

SAGE-Solexa versus ChIP-Seq-Solexa. Here we compare a transcriptomic assay versus an epigenomic assay performed with the same sequencing technology. One yields mRNA tags and the other genomic tags. As in the previous comparison, the mapping capacities are similar, although slightly better for ChIP-Seq. Here again, the percentage of uniquely mapped tags is maximum at 19 bp, 44% for LongSAGE versus 47% for ChIP-Seq, and then decreases until the maximum length (3.C versus 3.D). However, even at length 30 for ChIP-Seq, not all tags map a single genomic location, showing that the increase in length induces a loss of mapping capacity that is not compensated by an increase in unambiguous mapped locations (the dark blue bar does not fill the light blue one). Another phenomenon is involved with respect to the choice of an optimal tag length. Recall from 'Sequence errors' section that the probability of an erroneous nucleotide increases with its position in the tag. Thus, these considerations could explain why the prediction of single genomic locations may be more efficient with shorter tag lengths (19 or 20 instead of 30 bp).

This comparison also delivers measurements of technological and biological phenomena. We measured the percentage of biologically valid tags that are not mapped on the genome $\mathcal{R}(t)$. For a ChIP-Seq assay, the main cause of non-localization is the presence of individual variation (mainly single nucleotide polymorphism in our case). However, SNPs, but also post-transcriptomic modifications (e.g. RNA editing) and LongSAGE artefacts (e.g. a tag overlapping two exons) also affect transcriptomic assays. Hence, at a certain length, say $t = 20$, $\mathcal{R}_{chip}(t)$ estimates the percentage of non-localization due to SNPs, while the difference $\mathcal{R}_{sage}(t) - \mathcal{R}_{chip}(t)$ measures that due to causes specific to LongSAGE or transcriptomic assays. We observe that at length 20 SNPs affect at most 4.6% of tags (*a priori* in both ChIP-Seq and LongSAGE), while an additional 9.3% of LongSAGE tags may not be localized due to biological causes (Supplementary Table 2). The latter figure is equivalent to the one previously published for Sanger-LongSAGE in (7). Moreover, the percentages of erroneous occurrences at length 20 between the two assays are not significantly different (8.84 ± 0.09 for ChIP-Seq versus 9.14 ± 0.61, cf. Table 1). This suggests that the reverse transcription step performed in LongSAGE generates very little error.

Annotation

Above, we sought for ways to optimize the prediction of genomic regions of interest for a collection of experimental tags. Here we evaluate this strategy in practice for the prediction of transcriptionally active regions (TARs) using digital transcriptomic data. In our private SAGE-Solexa library, we selected all tags whose # *occ.* > 1, chose a tag length of either $t = 16$, 20 or 21 bp, mapped them to the human genome and classified the subset of uniquely mapped tags (cf. 'Classification of transcriptomic Tags' section) to evaluate how many predicted TARs are located in genes (exons, introns), in ESTs, and in intergenic regions. Table 3 gives the absolute numbers and percentages for each category and compares them with those of a human tiling array dataset (11).

By comparing analyses performed with different lengths, we notice that with $t = 21$ compared with $t = 20$, even if the percentage remains identical, one loses some predictions in all categories, but not for inxonic tags (overlapping an intron exon boundary). Indeed, the probability of an overlap increases with longer tags. With shorter tags $t = 16$, the percentage of exonic TARs drops to 34.7 compared with 38.5 with $t = 20$, and their number is 4-fold smaller. At $t = 16$, most tags mapped to multiple genomic locations, while at $t = 20$, 16 K more tags can be unambiguously associated with an exon. This illustrates the practical inconvenience of non-optimal lengths.

Clearly, with an appropriate length of 20 bp, the proportion of TARs in exonic and intergenic categories are similar between the tiling and digital transcriptomics analyses: 36% versus 38% of exonic TARs, and 29% (5.5 + 23.5) versus 29.5% (10.8 + 18.7) of intergenic TARs. The 38% of exonic tags/TARs represent 75% of the occurrences of uniquely mapped tags. As the classification schemes differ, the figures of intronic TARs cannot be directly compared. This confirms the validity of the prediction based on digital tags compared with hybridization-based data. Our analysis based on a single library suggests that a large proportion of TARs remain uncharacterized, i.e. belong to the so-called *dark matter* of the genome (20): 45% (23.5 + 6.6 + 8.8 + 0.3 + 5.5) if one considers an overlap with an EST as insufficient for annotation. This makes 25 240 tags located in intergenic regions or in antisense of annotated genes.

DISCUSSION

Direct ultra high-throughput sequencing is being applied to interrogate on a genome-wide scale, the transcriptome or the interactome with high specificity, sensitivity and statistical significance (21,22). Transcriptomic sequence census assays are paving the way to a thorough investigation and characterization of pervasive transcription and its annotation on the genome sequence. Although the standard annotation involves searching tags against transcript reference databases like UniGene or RefSeq, it has appeared that such gene-centred repositories fail to describe the whole complexity of mammalian transcriptomes (3,17). In mice or humans, significant genomic fractions that remain unannotated or are unlikely to code

for proteins give rise to RNA (1,8). For an in-depth investigation of the transcriptome, the tags produced must be mapped back on the genome, in reference to which annotations are accumulating (6). Once the genomic region of origin of a transcript has been determined, its characterization can build on other available annotations in its genomic neighbourhood. However, with the new generation of DNA sequencing systems, millions of data will be obtained to study the full scope of transcriptomes and genomes, but could also generate more and more erroneous data. It is crucial to evaluate the accuracy of these sequences. Here, we performed a global evaluation of DGE and ChIP-Seq experiments and proposed a new strategy for managing sequenced tags. Except where otherwise stated, all figures given below are for a 20 bp tag length.

Sequence errors

For the Sanger method, estimations of the sequence error level based on experimental measurements or on analysis of the PHRED score yielded rates <1% (17,18). Recently, an estimation was reported in the framework of genome resequencing with HTS techniques (23). Here we present an original and accurate method to estimate the sequence error for both DGE and ChIP-Seq assays. Our estimate of 0.57% erroneous nucleotides for tags sequenced with the Sanger method (Table 1) confirms published rates (17,18), and thereby validates our approach. A first interesting observation: the Solexa technique yields a similar, slightly lower error rate than that of Sanger (0.48 versus 0.57), showing that this technology is accurate enough for DGE and ChIP-Seq assays. The higher error level in LongSAGE data may be due to the diversity of platforms used to obtain SAGE libraries and to the error rate in the PCR step of that protocol (24).

With HTS, it has been suspected that the probability of an erroneous nucleotide increases with the sequence position (i.e. a higher error rate at the 3′-end) (25). Evidence of this bias was published for genome resequencing with the Solexa platform (23). The evidence provided here confirmed this hypothesis independently of the assay type. For DGE experiments, we only observed a weak bias between 14 bp and 20 bp, while a significant increase from 0.44 to 0.61 was noted between 20 and 30 bp in the ChIP-Seq data.

Compared with other methods, our approach enabled us to measure the impact of error on the prediction capacity. At a length of 19 bp, we found ≃36% of erroneous tags in both Solexa- LongSAGE and ChIP-Seq, and this rate increases with the tag length (Supplementary Table 3). This is the major cause of unmapped tags in both experiments. However, when selecting tags whose $\#\ occ. > 1$, the rate of erroneous tags drops <15%, thereby providing a clear rationale for implementing a widespread filtration criterion (6). Nonetheless, above this threshold, some erroneous tags remain, particularly in transcriptomic data due to the wide distribution of occurrence numbers. Indeed, highly abundant tags are more often sequenced and generate more erroneous occurrences, which mostly differ by one mismatch from the original tag. The latter can be identified by tag comparison and removed, as already proposed to further improve the data reliability (26). For example, in our Solexa library, a 20 bp tag with 8 165 occurrences generates a neighbourhood of 47 other tags, among which 44 have $1 < \#\ occ. < 20$, but none matches the genome.

Importantly, we have shown that above $\#\ occ. > 1$, the rate of false positive locations is very low ≃0.58% (value of $\mathcal{V}(20)_{\#\ occ.>1}$, Table 2). This is critical since Khattra *et al.* (24) experimentally demonstrated that rare tags also correspond with mRNA present in the sample, and could thus be biologically relevant. Our results also indicated that with a tag length like 20 (in the framework of a human study), which minimizes the background probability of being mapped and optimizes the prediction capacity, the fact that tags are mapped should help to distinguish erroneous tags from valid ones.

Causes of unmapped tags

As mentioned above, for both transcriptomic and ChIP-Seq data, sequence errors are the main cause of unmapped tags, but can be treated by filtering on the occurrence number and by choosing an adequate tag length (20 bp). Other causes may be artefacts due to the protocol or biological factors depending on the experiment, thus our comparison of DGE and ChIP-Seq data provides valuable information on both aspects. All figures below are for the complete tag set ($\#\ occ. > 0$).

For ChIP-Seq data, >62% of the tags can be mapped on the genome (Supplementary Table 2), while unmapped tags should be explained by sequence errors or polymorphism. We estimated that 4.6% of tags are affected by a SNP (Supplementary Table 2). Already published rates for LongSAGE were computed by multiplying the frequency of known SNPs on the genome by the total amount of sequence in a tag library, or by mapping mRNAs and tags on genomic SNPs. In both cases, these estimates, i.e. either 2% of tags with a frequency of 1/1000 (7) or 8.6% from the database of 2020 SNP associated tags (27), depend on the former state of the SNP collections. Moreover, both approaches neglect the fact that SNP collections pool the polymorphisms sampled in numerous individuals, while an RNA sample is taken from a single individual. Here our estimate is based on a single DNA library and on a single genome sequence, on much larger datasets, and is independent of SNP databases, which may provide a more realistic rate. However, these approaches do not account for individual copy number variations (28).

Transcriptomic assays yield more unmapped tags than ChIP-Seq ones, even when biologically unsure tags are removed, but they are hampered by the same error rate, which depends solely on the sequencing technique. Assuming an identical SNP rate for both types of data, we evaluated the difference at 9.4% of the tags ('Comparision of the mapping capacity of assays and sequencing techniques' section). These mapping failures could be attributed either to artefactual causes, e.g. tags overlapping an exon boundary or a poly-A tail (3), or to biological ones such as RNA editing or transplicing transcript (7,29). Our global rate of 9.4% closely agrees with

published rates: 9.6% for LongSAGE (7). For first time, our analysis generates an explanation for the fate of all unmapped tags.

The longest is not the best

The original SAGE protocol uses 14-bp tags. The goal of annotating the human genome with such transcript signatures (i.e. to tag both the transcript and its genomic region of origin) led to an extension of the tag length to 21 bp (6). The current tendency in sequence-based assays is to further increase the tag length, e.g. 36 and soon 50 bp with the Solexa/Illumina® technology, to potentially cover a larger diversity of tags. Longer tags generally imply using approximate rather than exact mapping. On the contrary, our analysis suggests that the longest is not the best. Indeed, longer tags have a much higher probability of including erroneous bases (0.48 with 20 bp versus 0.61 at 30 bp), and we see no reason why this should improve at 36 bp. They also more likely overlap an exon boundary or are affected by an SNP: two other causes of unmapped tags. Moreover, at 20 bp the chance of matching a genome location at random is already low, as is the rate of false positive locations, while the percentage of mapped tags deteriorates with length. Thus, exact mapping with an appropriate tag length provides a simple and secure method to process DGE and ChIP-Seq data. The MPSCAN program proves to achieve this task with perfect accuracy and high efficiency (13). Of course, this strategy is not restricted to using a prefix of the tag, but could easily be adapted to take a suffix, or a subpart of tags, as illustrated by the case of CAGE data ('Comparison of the mapping capacity of assays and sequencing techniques' section).

An interesting question is whether approximate instead of exact tag mapping would increase the percentage of uniquely mapped tags, especially for longer tags. For a given tag length, authorizing a few mismatches between the tag and the genome will automatically increase the number of mapped tags compared with exact mapping, but also that of multiply mapped tags and of false positives (since the background probability is higher). We studied this issue on a 34-bp ChIP-Seq dataset (GEO GSM325934). At length 34, ELAND, which authorizes up to two mismatches, yields the same number of uniquely mapped tags than MPSCAN does when mapping exactly the 20-bp tag prefixes (13). Hence, exact mapping with an appropriate tag length may yield more uniquely mapped tags than approximate mapping, and this is also valid with programs other than ELAND.

Although in the literature tag analyses generally exclude multimapped tags, as done in this work, one may ask whether longer tags could help to resolve multi-mapped tags. For this sake, we analysed the subset of multi-mapped ChIP-seq tags at a length t and their number of genomic locations at length $t + 2$, for $t = 16$ until 28 (Supplementary Table 1). For $t > 20$, the rate of resolved multi-mapped tags falls at 5%, and at length $t + 2$ the relative gain of uniquely mapped tags is completely annihilated by the number of tags that are not located anymore. Clearly, longer tags are not sufficient to rescue multi-mapped tags, but more complex computational strategies for this task represent a future line of research (30).

Detection of novel transcripts: towards an optimal strategy

High-throughput identification of transcribed genomic regions was recently performed with whole genome tiling arrays, which led to the discovery of pervasive transcription: numerous non-coding regions are transcribed (1,8). However, tiling arrays have two drawbacks: they cover only non-repetitive parts of the genome for design issues (1), and because of noise filtration it is necessary to focus on highly transcribed regions [e.g. the top 90th percentile in (1)]. But the accumulation of EST, MPSS and SAGE libraries over the years implies that the vast majority of yet unknown transcripts are low-abundance RNAs. Contrary to tiling arrays, DGE assays, like PMAGE, allow sampling at unprecedented depth RNAs transcribed anywhere on the genome. The tag sets they produce contain rare and biologically valid tags, but it is still a challenge to distinguish these from artefacts. We have shown that filtering tags whose # *occ.* = 1 eliminates most erroneous occurrences and tags, and that exact mapping of remaining tags with an appropriate length produces <0.6% false positive hits. Thus, our strategy enables the user to exploit rare tags, i.e. with a low # *occ.*, that are mapped on the genome. Hence, DGE, as an open technology, represents along with RNA-Seq (31), one of the most appropriate solutions to fully explore mammalian transcriptomes. The percentages of annotation in each category obtained by mapping the 20-bp prefix of our SAGE-Solexa library (Table 3) are closely in line with that of a tiling array study from the ENCODE project (11) and support this claim.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Skuld-Tech for the SAGE-Solexa library and for help in data processing, as well as Sophie Schbath for advice on word statistics. We acknowledge the support of the Languedoc-Roussillon Plateform MontpellierGenomix and the ATGC plateform at LIRMM.

FUNDING

French 'Ministère de l'Enseignement supérieur et de la Recherche' (PhD fellowship to N.P.); 'La ligue régionale contre le Cancer' Languedoc Roussillon; BioMIPS grants of the 'Université de Montpellier 2'. Funding for open access charge: ANR Project CoCoGen (BLAN07-1_185484).

Conflict of interest statement. None declared.

REFERENCES

1. Bertone,P., Stolc,V., Royce,T., Rozowsky,J., Urban,A., Zhu,X., Rinn,J., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
2. Kim,J., Porreca,G., Song,L., Greenway,S., Gorham,J., Church,G., Seidman,C. and Seidman,J. (2007) Polony Multiplex Analysis of Gene Expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, **316**, 1481–1484.
3. Rivals,E., Boureux,A., Lejeune,M., Ottones,F., Perez,O., Tarhio,J., Pierrat,F., Ruffle,F., Commes,T. and Marti,J. (2007) Transcriptome annotation using Tandem SAGE Tags. *Nucleic Acids Res.*, **35**, e108.
4. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
5. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
6. Saha,S., Sparks,A., Rago,C., Akmaev,V., Wang,C., Vogelstein,B., Kinzler,K. and Velculescu,V. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, **20**, 508–512.
7. Keime,C., Semon,M., Mouchiroud,D., Duret,L. and Gandrillon,O. (2007) Unexpected observations after mapping LongSAGE tags to the human genome. *BMC Bioinformatics*, **8**, 154.
8. Kapranov,P., Cheng,J., Dike,S., Nix,D., Duttagupta,R., Willingham,A., Stadler,P., Hertel,J., Hackermuller,J., Hofacker,I. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
9. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
10. Johnson,D.S., Mortazavi,A., Myers,R. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 149–156.
11. The ENCODE Project Consortium (2007) The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res.*, **17**, 732–745.
12. Kawaji,H., Kasukawa,T., Fukuda,S., Katayama,S., Kai,C., Kawai,J., Carninci,P. and Hayashizaki,Y. (2006) CAGE basic/analysis databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res.*, **34**, 632–636.
13. Rivals,E., Salmela,L., Kalsi,P., Kiiskinen,P., Tarhio,J. (2009) MPSCAN: fast localisation of multiple reads in genomes 9th Workshop on Algorithms in Bioinformatics (WABI'09) In: Steven S. Salzberg, T. Warnow (eds). *Lecture Notes in Bioinformatics*, Springer, Philadelphia, USA.
14. Robin,S., Rodolphe,F. and Schbath,S. (2005) *DNA, Words and Models*. Cambridge University Press, pp. 57–118.
15. Rivals,E. and Rahmann,S. (2003) Combinatorics of periods in strings. *J. Comb. Theory A*, **104**, 95–113.
16. Efron,B. and Gong,G. (1983) A leisurely look at the bootstrap, the Jackknife, and cross-validation. *Am. Stat.*, **37**, 36–48.
17. Piquemal,D., Commes,T., Manchon,L., Lejeune,M., Ferraz,C., Pugnère,D., Demaille,J., Elalouf,J.-M. and Marti,J. (2002) Transcriptome analysis of monocytic leukemia cell differentiation. *Genomics*, **80**, 361–371.
18. Colinge,J. and Feger,G. (2001) Detecting the impact of sequencing errors on SAGE data. *Bioinformatics*, **17**, 840–842.
19. Harbers,M. and Carninci,P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods*, **2**, 495–502.
20. Johnson,J.M., Edwards,S., Shoemaker,D. and Schadt,E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.
21. Velculescu,V.E. and Kinzier,K.W. (2007) Gene expression analysis goes digital. *Nat. Biotechnol.*, **25**, 878–880.
22. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
23. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
24. Khattra,J., Delaney,A.D., Zhao,Y., Siddiqui,A., Asano,J., Mcdonald,H., Pandoh,P., Dhalla,N., Prabhu,A.-L., Ma,K. *et al.* (2007) Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Res.*, **17**, 108–116.
25. Kharchenko,P., Tolstorukov,M. and Park,P. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
26. Akmaev,V.R. and Wang,C.J. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, **20**, 1254–1263.
27. Silva,A., Souza,J.D., Galante,P., Riggins,G., Souza,S.D. and Camargo,A. (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res.*, **32**, 6104–6110.
28. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
29. Li,X., Zhao,L., Jiang,H. and Wang,W. (2009) Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.*, **68**, 56–65.
30. Faulkner,G., Forrest,A., Chalk,A., Schroder,K., Hayashizaki,Y., Carninci,P., Hume,D. and Grimmond,S. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, **91**, 281–288.
31. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.