W378–W384 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 25 May 2010*
doi:10.1093/nar/gkq431

WebPrInSeS: automated full-length clone sequence identification and verification using high-throughput sequencing data

Andreas Massouras[1], Frederik Decouttere[2], Korneel Hens[1] and Bart Deplancke[1,*]

[1]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland and [2]Genohm BVBA, B-9052, Zwijnaarde, Belgium

Received February 20, 2010; Revised April 30, 2010; Accepted May 6, 2010

ABSTRACT

High-throughput sequencing (HTS) is revolutionizing our ability to obtain cheap, fast and reliable sequence information. Many experimental approaches are expected to benefit from the incorporation of such sequencing features in their pipeline. Consequently, software tools that facilitate such an incorporation should be of great interest. In this context, we developed WebPrInSeS, a web server tool allowing automated full-length clone sequence identification and verification using HTS data. WebPrInSeS encompasses two separate software applications. The first is WebPrInSeS-C which performs automated sequence verification of user-defined open-reading frame (ORF) clone libraries. The second is WebPrInSeS-E, which identifies positive hits in cDNA or ORF-based library screening experiments such as yeast one- or two-hybrid assays. Both tools perform *de novo* assembly using HTS data from any of the three major sequencing platforms. Thus, WebPrInSeS provides a highly integrated, cost-effective and efficient way to sequence-verify or identify clones of interest. WebPrInSeS is available at http://webprinses.epfl.ch/ and is open to all users.

INTRODUCTION

Progress on our understanding of how biological processes operate at a systems level has been in part propelled by the availability of 'omics' approaches that allow the probing of gene or protein function at a large or genome-wide scale. Critical for the development of many of these approaches has been the improvement in genome annotation, which allowed the generation of comprehensive sets of protein-encoding open reading frames (ORFs) in versatile cloning format (1). Consequently, the same ORF clones can be used for a wide range of approaches that require exogenous protein expression. Examples include protein arrays on which proteins are derived from cDNA or ORF templates (2,3), systematic cellular protein overexpression projects (4) or high-throughput yeast one- or two-hybrid protein-DNA or protein-protein interaction mapping assays (5,6). Given the experimental value of versatile ORF clones, many small-to-large ORF cloning projects have been initiated across a wide range of organisms such as viruses, bacteria, plants and animals (7). To use these ORF clones as reliable templates for protein expression, each clone should ideally be fully sequenced and compared to the reference ORF sequence to evaluate whether the ORF clone can be accepted for protein synthesis. Full-length ORF sequencing is however costly, and aligning and verifying each sequence to its reference counterpart is cumbersome. Valuable efforts to automate this ORF clone evaluation procedure based on conventional Sanger reads have been undertaken (8), however, the considerable costs associated with large-scale Sanger sequencing will continue to impede extensive sequence verification efforts. It is therefore not surprising that such evaluation procedure has been performed for only a limited number of clone collections (8).

In recent years, high-throughput sequencing (HTS) technologies such as Solexa (Illumina), SOLiD (Life Technologies) or 454 (Roche) have become available that are revolutionizing our ability to obtain cheap, fast and reliable sequence information (9). These features make HTS an excellent solution for overcoming the problems involved in sequence verification of large clone collections described earlier. Indeed, rather than using conventional Sanger sequencing, it may be less costly

*To whom correspondence should be addressed. Tel: +41 21 693 1821; Fax: +41 21 693 0980; Email: bart.deplancke@epfl.ch

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

and require less effort to sequence-verify in batch a set of ORF clones using HTS. Such approach would already become cost-effective from just 50–60 clones depending on average ORF length. In addition, providing an output of 10 million 36 bp reads for one Solexa lane, a library containing ~1450 5 kb clones could be fully sequenced at 50× coverage at minimal cost. With an ever increasing number and length of reads at an ever decreasing sequencing cost (9), the number of clones that can be fully validated in a single HTS run will continue to rise. While the experimental part may be straightforward, the downstream data analysis is less simple and prevents this sequence verification approach from becoming a routine lab technique. Several bioinformatics tools are available that allow the identification of sequence variants by aligning reads to a reference sequence [e.g. (10,11)]. However, these tools require advanced computing skills and they also do not allow the characterization of variants that are >3–5 bp (12). Moreover, most structural variation mapping tools do not support sequence assembly and thus the precise variant sequence cannot be retrieved (12). The latter is important as it provides information as to the precise consequence of the different sequence variants on protein expression (e.g. a silent mutation, introduction of a stop codon, disruption of a functional domain etc.) (8), and may even lead to the identification of a known or novel splice form (13). We developed a new tool, Primer-Initiated Sequence Synthesis of Clones (PrInSeS-C), which allows the bi-directional assembly of ORF clone sequences that were analyzed in batch using HTS based on a novel primer-initiated assembly procedure introduced by Massouras *et al.* (14), and explained in more detail in Supplementary Figure S1a. Subsequently, we paired PrInSeS-C to an automatic decision making workflow that we developed, which, together with an easy-to-use web interface (WebPrInSeS-C), should now enable even non-initiated users to fully sequence-verify their clone collections in a simple and cheap fashion using HTS.

The ability to identify and assemble clone sequences from HTS data has several other applications. One of these is the identification and allelic characterization of cDNA (or ORF) clones that were found as positives in cDNA (or ORF) library-based protein, drug or DNA interaction screens. As these assays often produce several hundred positive hits, interactor identification is costly and therefore typically limited to one-ended Sanger sequencing, generating an interaction sequence tag (IST)(15–17). Since these ISTs tend to cover only part of the cDNA, critical allelic information (such as different splice forms, single nucleotide polymorphisms (SNPs), truncations), which provides important detail on the nature of the interaction, is often lost. Again, HTS has the capacity to not only identify but to also full-length sequence positive library hits in a single run. Moreover, the use of HTS bar codes may even allow the simultaneous processing of positives from distinct library screens, further reducing cost. We therefore developed a second easy-to-access web tool, WebPrInSeS-E, to identify and assemble clone sequences derived from cDNA or ORF library screens based on the same novel primer-based assembly procedure introduced by Massouras *et al.* (14), and discussed in more detail in Supplementary Figure S1b.

WebPrInSeS

Input file description

Part of the WebPrInSeS submission page is shown in Figure 1. Under the Tools tab, the user has the option to select either WebPrInSeS-C (for sequence verification) or WebPrInSeS-E (for the identification of positives derived from cDNA or ORF library screens).

WebPrInSeS-C. To use WebPrInSeS-C, the user needs to upload three separate files using the 'Upload File' menu on the left (Figure 1). The first is a fasta file which contains the ORF reference sequences exactly as they are inserted in the respective vector (this will typically be derived from preparatory sequence characterization of each individual clone). For example, if a stop codon is not contained within the ORF sequence (a so-called 'open format' ORF clone which allows C-terminal tagging of the corresponding protein), then it should not appear in the reference sequence either. The reference sequences will be used to locate variants in the experimental sequences of interest.

Prior to sequencing a batch of clones, it is beneficial to analyze whether the selected sequencing technology will support the complete assembly of the sequences of interest. This is because a region that appears twice or more within the same clone or in other clones and that is longer than the acquired read length, will not be resolved during the assembly process. The longer the read length, therefore, the less duplicate regions will constitute a problem for sequence assembly. Consequently, 454 sequencing, which produces reads up to 500 bp, would be a logical choice for clone sequence verification. However, the number of reads produced by a 454 versus, for example, a Solexa machine is significantly lower, thus restricting the number of clones that can be analyzed simultaneously. Also, the user may not have the option of choosing between different HTS technologies, and thus we added a small tool, Duplicate Finder, that allows the user to identify potentially problematic clones based on the length of the generated reads. This tool allows the user to enter a specific read length, and uses as input the reference sequence fasta file. After running the tool, the names of problematic clones or sequences are displayed, as well as the number and length of the duplications. Depending on the results, the user may opt to change the selected read length (e.g. Illumina currently offers the possibility of producing reads from 35 up to 100 bp with the latter more expensive than the former), or may decide to exclude certain problematic clones from the analysis and use conventional Sanger sequencing for those. A Duplicate Finder submission page screenshot as well as an output example are shown in Supplementary Figure S2 and Table S1.

A second input that WebPrInSeS-C requires is a file, in fasta format, containing the short sequences (typically one

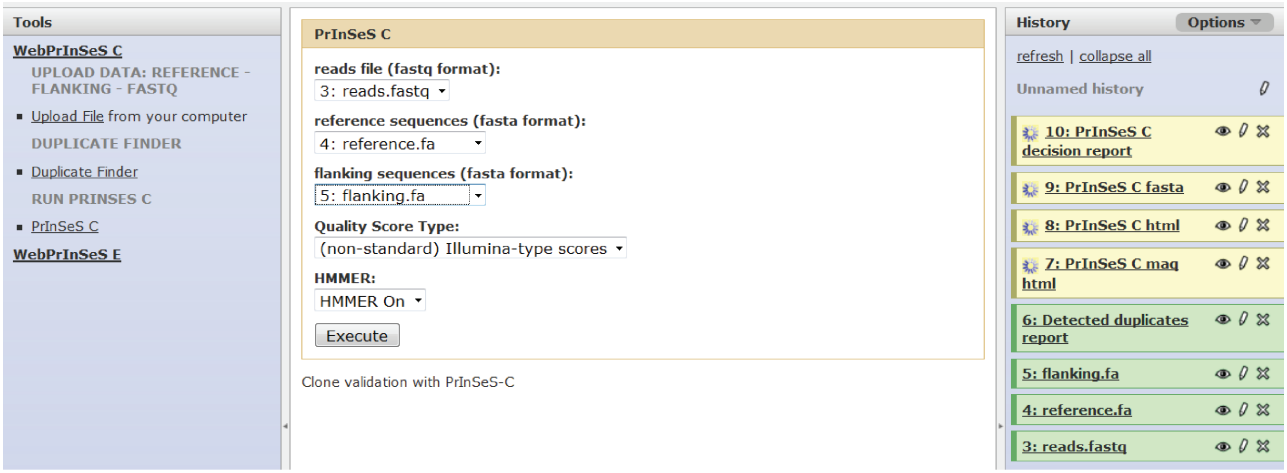

Figure 1. WebPrInSeS-C Screenshot of the web interface after successfully launching WebPrInSeS-C. WebPrInSeS-C requires a reads file (reads.fastq in the example), a fasta file containing the flanking sequences (flanking.fa in the example) and fasta file containing the DNA reference sequences (reference.fa in the example) to start assembly. It outputs a tab-separated file containing summary information of the processing done by Maq and PrInSeS-C (PrInSeS-C decision report), a fasta file with the assembled sequences (PrInSeS-C fasta), an html file which visualizes sequences of interest as assembled by PrInSeS-C and aligned back to the reference (PrInSeS-C html) and a file containing the results of the Maq alignment and the subsequent data processing (PrInSeS-C Maq html).

to two times the read length) adjacent to the sequences of interest in the vector (the so-called flanking sequences). This file should therefore contain two sequences: (i) a 5′ flanking sequence in reverse complement format, and (ii) the 3′ flanking sequence. These flanking sequences are required as WebPrInSeS-C uses them as terminators of the assembly process, so that each sequence of interest will be assembled in both directions (hence the reverse complement requirement for the 5′ flanking sequence). If these sequences are not known by the user, then these should be determined using conventional Sanger sequencing prior to the analysis. It should be noted that if distinct vectors were used within the same batch of clones, than the user should run WebPrInSeS-C as many times as the number of vectors used, while each time changing the content of the flanking sequence file and the corresponding reference sequence file, but using the same reads file. Importantly, the total execution time will not be significantly affected as in this way the user is partitioning the data into smaller sets.

The final input that WebPrInSeS-C requires is the reads file in the widely used and convenient fastq format with either Illumina or Phred-type scores. If there are several reads files, they should be concatenated into one. However, reads from experiments in which multiple clone sets were sequenced simultaneously by using bar codes should first be clustered per bar code in separate files, after which the respective bar codes should be removed and the resulting files should be analyzed separately. As the reads file is much larger than the reference and flanking sequence files, it will take longer to upload. We have tested the upload speed from outside our internal network (from Ghent University, Belgium to EPFL, Switzerland) and found that it takes about 5–7 min/Gb of data to upload over the internet. Upload time will however largely depend on the speed of the user's internet connection.

WebPrInSeS-E. For WebPrInSeS-E, a similar procedure as for WebPrInSeS-C must be followed. The user must first upload a file with the vector flanking sequences used for the library screen as well as the reads file. In contrast to WebPrInSeS-C, if several vectors were used for the library screens, then all flanking sequences may be included in a single fasta file, but again listing each 5′ flanking sequence in reverse complement format. A reference sequence file is not required since the identity of the sequences of interest is not known.

Workflow

WebPrInSeS-C. To launch the program, the user needs to click on the 'PrInSeS-C' link under the WebPrInSeS-C header. The reference and the flanking sequence as well as the reads files must be selected from the corresponding pick-lists (Figure 1). The user must specify if the reads file contains Illumina or standard quality scores. In addition, the user has the option to run the Hmmer tool with global pfam profiles (18) in order to identify the functional domains contained in the reference sequences. Since running Hmmer substantially increases the execution time, we have not included it as a default setting. However, running Hmmer provides an additional sequence verification step as it allows our decision tool to analyze whether there are any variants within known functional domains, which may lead to an automatic rejection of the respective clone depending on the nature of the variant (see below). Clicking on 'Execute' then launches the WebPrInSeS-C pipeline. The user can return to the same page at a later time to view and/or download the results, which appear on the right (Figure 1).

The assembly and decision workflow is shown in respectively Figure 2a and b. Reads are both aligned to the reference sequences using the read mapper Maq (11) and assembled and compared to the reference sequences

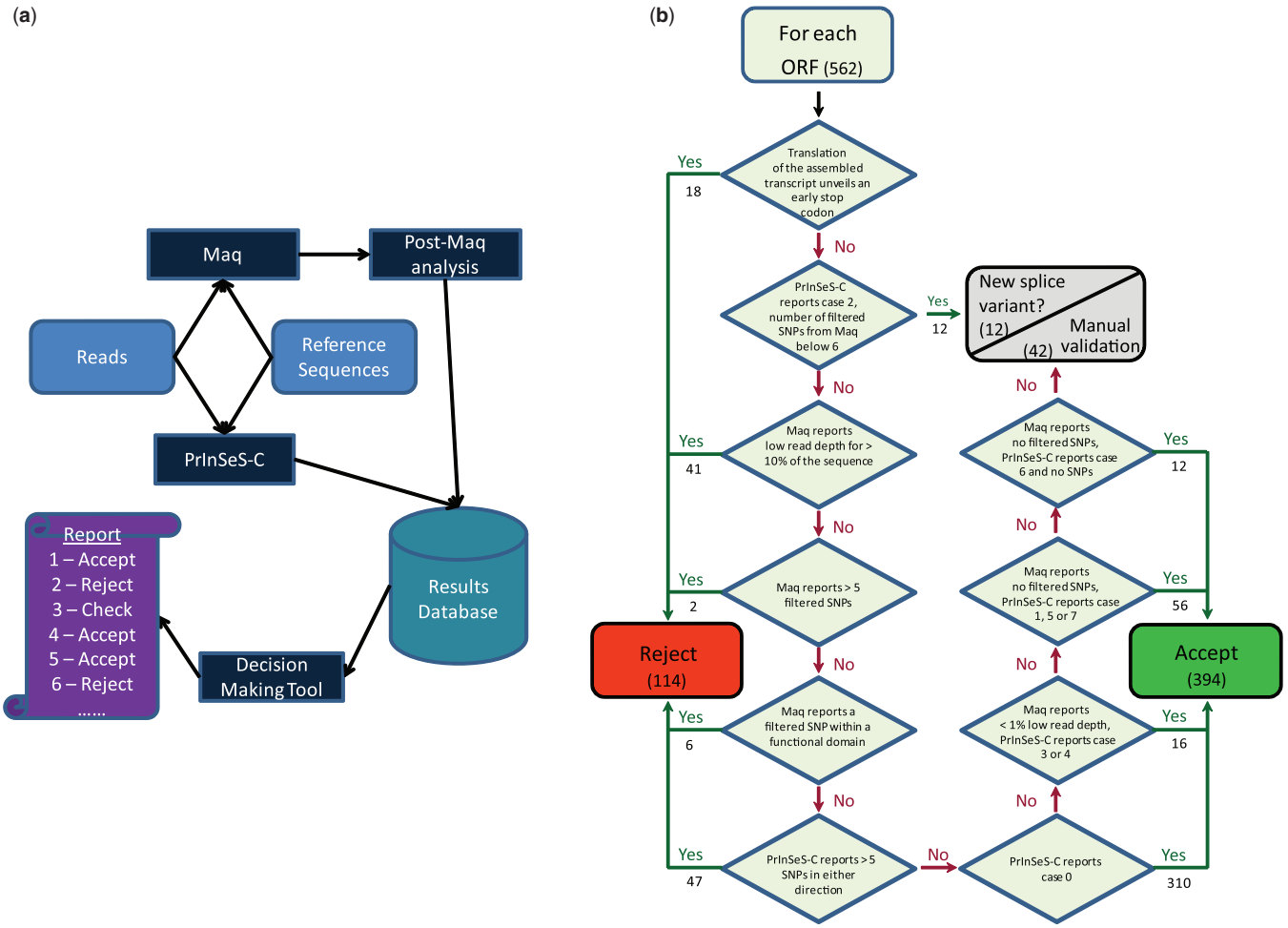

Figure 2. The automated clone validation pipeline. (**a**) Diagram outlining the workflow. (**b**) The heuristic algorithm of the decision-making tool. Numbers on the branches indicate the number of clones from a sequenced *Drosophila* ORF collection that fall in each category (see 'A working example' section).

using PrInSeS-C. The results from both analyses are stored in a relational database, which constitutes the foundation for the implementation of a set of heuristic rules leading to the acceptance or rejection of the clones. Specifically, clones are evaluated using an automated pipeline that takes into account several factors: the first is the way a sequence of interest has been assembled. We consider nine distinct assembly scenarios (see also Supplementary Figure S3): Cases 0, 1 and 2 indicate full assembly in both directions but respectively without or with mismatches or insertions/deletions relative to the reference sequence; Case 3 reflects partial assembly on one strand; Cases 4 and 5 indicate partial assembly on both strands but respectively match or deviate from the reference sequence when combined; Case 6 indicates an early assembly stop on both strands without overlap, whereas Cases 7 and 8 indicate that strands overlap but are distinct with respectively only one matching the reference sequence or both not matching but in a different fashion. A second evaluation factor is the number of mismatches (or SNPs) detected by the Maq and PrInSeS-C programs and a third factor is the number of areas that exhibit low read depth (Figure 2b). Additionally, SNPs are classified according to their likely impact on protein function. For this, it is required to use the Hmmer tool to first identify the functional domains within the selected sequences of interest. SNPs causing synonymous or conservative mutations are ignored. However, if there are more than five remaining or 'filtered' SNPs resulting in non-conserved amino acid changes in the sequence of interest or if there is one such SNP in a functional protein domain, than the respective clone is rejected. Clones containing a SNP that produces an early stop codon will also be rejected. All the factors are evaluated in the given order (Figure 2b); if any condition is met, a decision is made and no further conditions are considered.

WebPrInSeS-E. To launch the program, the user needs to first select the file containing the flanking sequences as well as the reads file from the corresponding pick-lists (Supplementary Figure S4). In addition, the user must select the organism from which the cDNA or ORF library was derived from the indicated pick-list. This is because WebPrInSeS-E uses all reported exons for this organism longer than the read length (based on the

latest Ensembl version) (i) as Maq alignment templates to identify the positive hits based on read coverage (any exon exhibiting a read depth >10 is considered as being present) and (ii) as possible primers for the assembly process. The latter assures the full-length allelic characterization of positive hits derived from library screens providing that these hits have sufficient read coverage throughout the cDNA or ORF. Currently, we have included the most common model organisms (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, zebrafish, mouse and human), but can expand this list upon request. Clicking on 'Execute' then launches the WebPrInSeS-E pipeline. Similar to the WebPrInSeS-C platform, the user can return to the same page at a later time to view and/or download the results, which appear on the right.

Output

WebPrInSeS-C. A run of WebPrInSeS-C processing ~1.6 GB of reads data takes ~70 min without invoking Hmmer and 5–6 h when a protein domain analysis is included. The output files, which appear in the right column of the execute page and which can be visualized by clicking on the 'eye' symbol (Figure 1) include (i) a 'PrInSeS-C Maq html' file containing the results of the Maq alignment and the subsequent data processing (Figure 3a), (ii) a 'PrInSeS-C html' file, which visualizes sequences of interest as assembled by PrInSeS and aligned back to the reference (Figure 3b), (iii) a 'PrInSeS-C fasta' file with the assembled sequences and finally (iv) 'PrInSeS-C decision report', a tab-separated file which can be opened using a spreadsheet application, containing summary information of the processing done by Maq (third to seventh column) and PrInSeS-C (from the eighth column onwards) (Supplementary Table S2). The second column contains the decision on the respective clone using the set of heuristic rules explained above and shown in Figure 2b. The ability to view the results in spreadsheet format enables the user to implement his/her own set of rules on accepting or rejecting clones. This can be easily done through the implementation of additional ranking or filtering formulae on the spreadsheet. For example, we are currently accepting clones that are only partially assembled (Case 6), but exhibit a 'clean' Maq profile. Users may however opt to only accept clones if they were fully assembled by PrInSeS-C. Nevertheless, we strongly encourage users to provide feedback on the most optimal set of acceptance or rejection criteria.

WebPrInSeS-E. A run of WebPrInSeS-E processing ~1.6 GB of data takes less than an hour. Output files are displayed in similar fashion as for WebPrInSeS-C. One file is 'Hit List' containing a list of identified positives displayed as Ensembl Gene IDs. In addition, next to each Gene ID is a number representing the average read depth across all identified exons linked to this gene. As this number correlates with how many times the respective positive hit is present in the positive hit library, it provides an overall confidence measure in the detected interaction. A second file is 'PrInSeS-E html' visualizing the assembled transcripts, each aligned to the reference sequence of the most similar, documented transcript (Supplementary Figure S5). A final file is the 'PrInSeS-E fasta' file with the sequences of the assembled transcripts themselves in fasta format.

A WORKING EXAMPLE

WebPrInSeS-C

We evaluated the performance of the WebPrInSeS-C work flow by processing read data derived from an in-house generated library containing 562 distinct *Drosophila melanogaster* ORF clones that was sequenced in one Solexa lane producing 76 bp reads (K. Hens et al., manuscript in preparation). As a proof-of-principle, we first ran Duplicate Finder and found eight combinations representing 15 problematic clones (Supplementary Table S1). After processing the 7 941 527 reads using the WebPrInSeS-C workflow, 394 ORF clones (70%) were accepted whereas 121 (20%) were rejected (Figure 2). The predominant causes of rejection were the detection by PrInSeS-C of more than five SNPs in both assembly directions, as well as the occurrence of low read depth regions (<10-fold coverage) in >10% of the assembled sequence. For the remaining 54 (10%), no automatic decision could be reached based on the decision algorithm and thus these clones were curated manually, resulting in the identification of 12 likely alternative splice forms as these sequences were found to contain in-frame insertions or deletions compared to the reference. Examples of the Maq and PrInSeS-C output for a manually curated and an accepted clone are shown in Figure 3.

WebPrInSeS-E

We used the same data set described above to evaluate the WebPrInSeS-E performance. After processing, WebPrInSeS-E identified 549 out of 562 (98%) possible clones. The remaining clones may not have been detected because of technical issues (low coverage) or simply because they were not included in the library due to mistaken identity. Importantly, for about half of these (265 or 48%), WebPrInSeS-E was able to generate a full sequence assembly. It should thereby be noted that data from screens using ORF- rather than cDNA-based libraries (as is theoretically the case here) should preferably be analyzed using WebPrInSeS-C if the sequence identity of the ORFs within the library is known. As is shown in the paragraph above, overall assembly performance by WebPrInSeS-C is greater in this case, as the presence of UTRs in the reference exons but not in the corresponding sequenced ORFs reduces the number of sequences that can be assembled by WebPrInSeS-E (Supplementary Figures S1 and S5).

IMPLEMENTATION

WebPrInSeS was implemented on top of the Galaxy bioinformatics framework (http://g2.bx.psu.edu/). The Galaxy server itself is written in Python and requires Python 2.4 or up to run, but is able to execute any kind

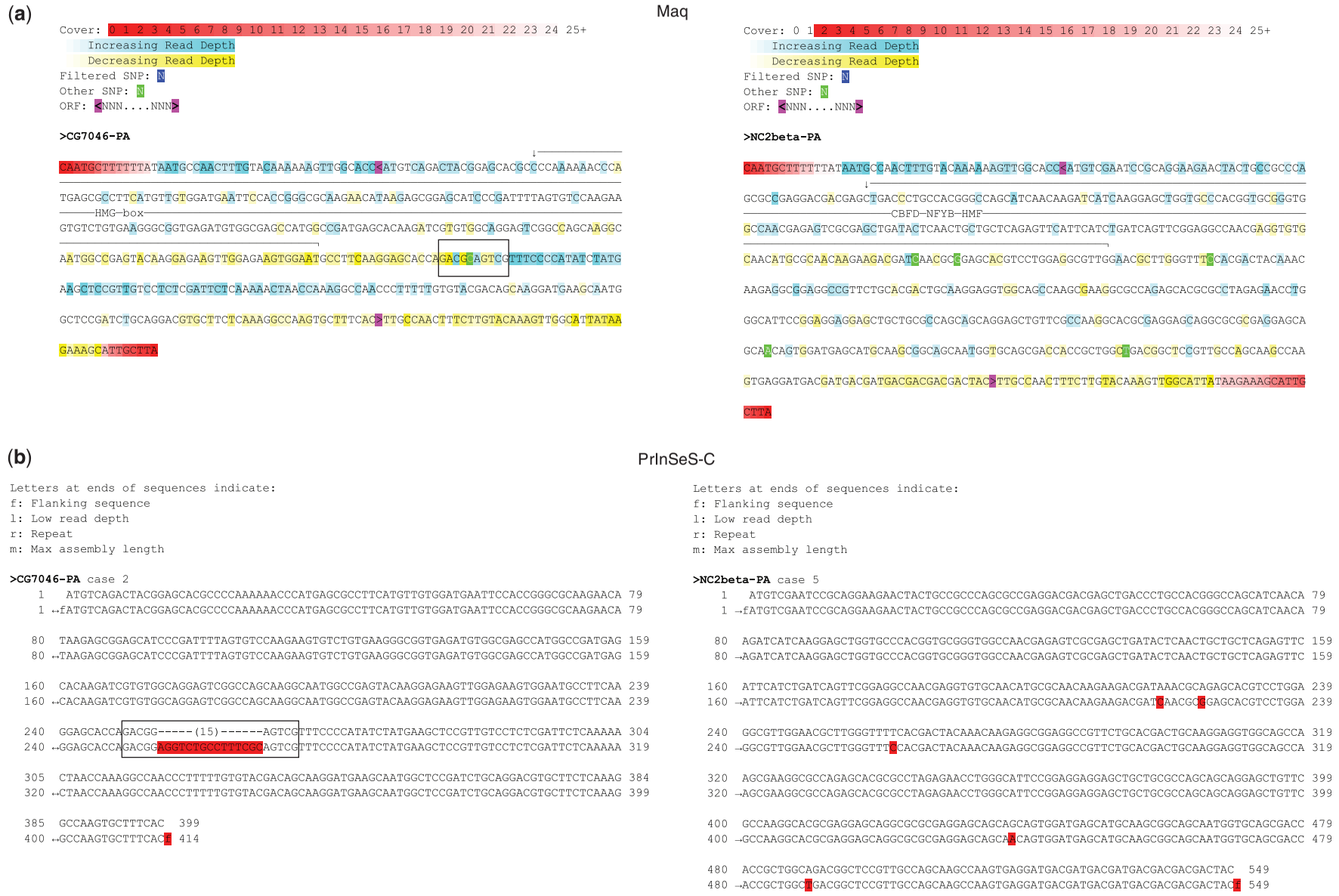

Figure 3. Visualization of the PrInSeS-C Maq html file and the PrInSeS-C html file. (Left) Example of a clone (CG7046-PA) that was referred for manual curation. In this case, it concerned an alternatively spliced transcript. The position where WebPrInSeS-C via assembly detected a 15-nt insertion is boxed. (Right) Example of an accepted clone (NC2beta-PA). (**a**) Visualization of the PrInSeS-C Maq html file for both examples. Yellow boxes mark a drop in read depth; cyan boxes mark recovery in read depth. Red colored sequences indicate regions of low read-depth from zero (dark red) to 25 (white)-fold coverage. Dark blue boxes highlight non-synonymous mutations, green boxes highlight synonymous mutations. Functional protein domains are highlighted by the presence of a line above the respective coding sequence as well as their ID. (**b**) Visualization of the PrInSeS-C html file for both sequences. Red boxes indicate SNPs or indels. The case number indicates a specific assembly scenario as described in the main text.

of bioinformatics tool without restrictions on execution language (e.g. Perl, Bash, Java, ...). WebPrInSeS provides the necessary shell scripts to execute the WebPrInSeS perl scripts from within Galaxy. Input (files) and output (results) parameters and file types are fully described in the Galaxy configuration files for WebPrInSeS, available for download at http://updepla1srv1.epfl.ch/galaxy/static/galaxy-webprinses-conf.tar.gz. Furthermore, if users wish to run PrInSeS-C or -E locally, they may download the source code along with the user manual and tutorials from http://prinses.epfl.ch.

CONCLUSION

The raw sequencing power of the new HTS technologies holds the promise of boosting experimental applications that require deep and inexpensive sequencing. The WebPrInSeS interface provides researchers with an easy-to-use tool to interpret the millions of short reads produced by these technologies for two such applications: WebPrInSeS-C provides a highly automated pipeline to sequence-verify ORF clone libraries in a single step, combining *de novo* assembly of the clone sequence with an in-house generated decision making workflow to evaluate the newly assembled sequence. WebPrInSeS-E identifies hits from library screening experiments, again performing *de novo* assembly where possible. Both applications of WebPrInSeS allow the user to unlock the power of HTS technologies for the straight-forward and cost-efficient identification and sequence verification of clones.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Carine Gubelmann, Christopher Chidley and Kai Johnsson for help with the testing of WebPrInSeS and all the members of the Deplancke Lab for critical reading of the manuscript.

FUNDING

Funding for open access charge: Swiss National Science Foundation; Swiss National Center of Competence in Research in Genetics; Marie Curie International Reintegration Grant (BD) from the Seventh Research Framework Programme; Institutional support from the Ecole Polytechnique Fédérale de Lausanne.

Conflict of interest statement. None declared.

REFERENCES

1. Rual,J.-F., Hill,D.E. and Vidal,M. (2004) ORFeome projects: gateway between genomics and omics. *Curr. Opin. Chem. Biol.*, **8**, 20–25.
2. Ramachandran,N., Raphael,J.V., Hainsworth,E., Demirkan,G., Fuentes,M.G., Rolfs,A., Hu,Y. and LaBaer,J. (2008) Next-generation high-density self-assembling functional protein arrays. *Nat. Methods*, **5**, 535–538.
3. Hu,S., Xie,Z., Onishi,A., Yu,X., Jiang,L., Lin,J., Rho,H.-s., Woodard,C., Wang,H., Jeong,J.-S. *et al.* (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell*, **139**, 610–622.
4. Theodorou,E., Dalembert,G., Heffelfinger,C., White,E., Weissman,S., Corcoran,L. and Snyder,M. (2009) A high throughput embryonic stem cell screen identifies Oct-2 as a bifunctional regulator of neuronal differentiation. *Genes Dev.*, **23**, 575–588.
5. Vermeirssen,V., Deplancke,B., Barrasa,M.I., Reece-Hoyes,J.S., Arda,H.E., Grove,C.A., Martinez,N.J., Sequerra,R., Doucette-Stamm,L., Brent,M.R. *et al.* (2007) Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat. Methods*, **4**, 659–664.
6. Rual,J.-F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
7. Yashiroda,Y., Matsuyama,A. and Yoshida,M. (2008) New insights into chemical biology from ORFeome libraries. *Curr. Opin. Chem. Biol.*, **12**, 55–59.
8. Taycher,E., Rolfs,A., Hu,Y., Zuo,D., Mohr,S., Williamson,J. and LaBaer,J. (2007) A novel approach to sequence validating protein expression clones with automated decision making. *BMC Bioinformatics*, **8**, 198.
9. Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
10. Li,H. and Durbin,R. (2009) Fast and accurate long read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
11. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
12. Medvedev,P., Stanciu,M. and Brudno,M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.
13. Salehi-Ashtiani,K., Yang,X., Derti,A., Tian,W., Hao,T., Lin,C., Makowski,K., Shen,L., Murray,R.R., Szeto,D. *et al.* (2008) Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nat. Methods*, **5**, 597–600.
14. Massouras,A., Hens,K., Gubelmann,C., Uplekar,S., Decouttere,F., Rougemont,J., Cole,S.T. and Deplancke,B. (2010) Primer-initiated sequence synthesis to detect and assemble structural variants. *Nat. Methods*, in press.
15. Koegl,M. and Uetz,P. (2007) Improving yeast two-hybrid screening systems. *Brief. Funct. Genomic. Proteomic.*, **6**, 302–312.
16. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
17. Deplancke,B., Mukhopadhyay,A., Ao,W., Elewa,A.M., Grove,C.A., Martinez,N.J., Sequerra,R., Doucette-Stamm,L., Reece-Hoyes,J.S., Hope,I.A. *et al.* (2006) A gene-centered C. elegans protein-DNA interaction network. *Cell*, **125**, 1193–1205.
18. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.