diagenode
A Hologic Company

APPLICATION NOTE

Successful application of novel plant ChIP-seq technology results in high quality sequencing data from *Arabidopsis thaliana* seedling cells

Miklós Laczik, Céline Sabatel, Mareike Hohenstatt, Jan Hendrickx, Dominique Poncelet

Introduction

In the recent years massively parallel sequencing (also known as Next Generation Sequencing or NGS) has become the dominant method for sequencing, and the decreasing costs and increasing yields transformed biosciences. Due to greater accessibility and higher throughput, NGS is also highly utilized in epigenetic or epigenomic studies, most widely in ChIP-seq. NGS is the detection method of choice for ChIP, which gives an accurate qualitative and quantitative insight into protein-DNA interactions, such as binding events and sites of transcription factors or chromatin modifications including methylation and acetylation.

In addition to the widespread use of ChIP-seq and other NGS-related techniques, another major trend is the increasing use of limited amounts of starting material (e.g. due to ancient fossil samples, forensic samples, or limitations on amounts of material generated cells). In some instances, samples such as may not yield enough material for a regular ChIP-seq as in the case of embryonic tissues, the few cell layers of a specific zone in the apical meristem, etc. In other studies the researcher might be simply interested in the epigenetic events in a single cell.

Plant researchers also have a need for ChIP-seq from limited amounts of material, but their needs are rarely met. There are not many commercial solutions available that are dedicated to plant research, and usually these are not designed for low cell number experiments.

In this study, we show a successful and reliable ChIP-seq technology using limited amounts of plant material, with specific reagents and protocols, supported by strict bioinformatic QC analysis.

Methods

Arabidopsis thaliana is one of the most studied model organism in the plant research community, having given valuable insights into plant genetics, epigenetics, development and evolution. Therefore we chose *Arabidopsis thaliana* (ecotype Columbia) seedlings for our ChIP-seq experiment. We performed two series of biological replicates, with different amounts of starting materials, and as a reference we used samples processed with the Illumina TruSeq kit, and we also compared our results to public dataset of ChIP-seq on A. thaliana seedlings. ChIP assays were performed using the Diagenode Plant ChIP kit (Cat. No. C01010150) on 0.25 g freshly weighed plant tissue with the antibody against the histone mark H3K4me3 (Cat. No. C15410003). Bioruptor Pico (Cat. No. B0160001), thermo-controlled sonication device, which was used for the shearing of the chromatin. The immunoprecipitated DNA was used for library preparation using the MicroPlex Library Preparation kit (Cat. No. C05010010) on different amounts of DNA (1 ng, 500 pg and 100 pg) and using the Illumina TruSeq kit on 5 ng of DNA. Sequencing was done on Illumina HiSeq 2000. The overview of the experimental design is in Table 1.

We use our proprietary bioinformatics pipeline to analyse the data and perform a strict QC. Some of the QC methods are derived from the ENCODE/modENCODE projects, which are considered gold standards for ChIP-seq. See the concepts of FRIP, NSC, RSC in the paper detailing the ENCODE/modENCODE guidelines (Landt et al: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31.) The overlap criteria are also defined by the ENCODE/modENCODE criteria, where at least 80% of the top 40% of the peaks of two replicates should have a match in the replicate dataset.

Results

After the ChIP and library preparation, samples were sequenced. The sequencing results were of very high quality (Figure 1).

We performed the alignment, peak calling and independent quality analysis of the samples. Both the ENCODE/modENCODE defined strand cross-correlation analysis, and the other quality indicators proved that all of the datasets are high quality. In many cases the samples prepared with our proprietary kit are higher quality than those prepared with the Illumina TruSeq kit or various publicly referenced protocols. From very low amounts of starting material the peak numbers, peak widths, and probability scores decrease slightly, which is expected. Refer to Table 2 for the quality indicators and Figure 2 for the visualization of the strand cross-correlation analysis.

Figure 3 compares the average peak coverage of a series. The bar charts show the average coverage along the length of all peaks, per percent of the length. It is visible that the peak shapes are very similar, and for all samples we achieved an even and sufficiently high coverage.

We also made several peak overlap studies to reveal the similarity between datasets, to learn how reproducible our results are, and to see if we detected the peaks that are in the reference datasets. In summary, we have outstanding reproducibility and similarity, as all of the datasets performed far above the ENCODE/modENCODE, defined at an 80% cutoff value (even a 100% overlap was not uncommon). See Table 3 for further details. Figure 4 shows a screenshot of perfectly overlapping peaks.

To be absolutely sure that the peaks indeed overlap each other, and there is no bias in the datasets (e.g. partial overlapping, shifted overlapping as in covering only one end of the reference peaks, etc.) we generated further evidence. Figure 5 shows the average overlap profile for one of the series, calculated by computing the shift between each overlapping pair (how much is a peak orientated left or right from the reference peak) and the coverage (how much of the whole length of the reference peak is covered). We observed a generally central overlap between the peaks with no bias.

Figure 6 is another representation of the shifting bias analysis, as the box-and-whiskers plots show the distribution of the shifts. Although there are some outliers (whiskers), the significant population of the shifts (the two inner quartiles, represented by the boxes) concentrate tightly and symmetrically around zero.

Discussion

We have shown that with our novel plant ChIP-seq kit, it is possible to gain high quality, consistent, and reproducible results, that highly correlate with reference datasets. Even with limited amounts of sample we reproduced the peaks of datasets derived from high sample amount experiments. Even those datasets that are not replicate show a level of similarity that is found between true replicates. We also proved that the datasets contain no bias, the peaks are evenly covered, and the overlapping pairs are not shifted.

In addition to our plant ChIP-seq kit, we also recommend our Bioruptor for shearing and our ChIP-seq grade antibodies, to achieve optimal ChIP-seq with plant samples.

Acknowledgements

We would like to thank Fasteris for their high quality Next Generation Sequencing service.

Table 1									
Code name in experiment	**651**	**652**	**653**	**654**	**661**	**662**	**663**	**664**	**REF**
Biological replicates	Replicate series 1				Replicate series 2				Reference dataset
Provider of library preparation kit	Diagenode			Illumina	Diagenode			Illumina	
Provider of ChIP-seq kit	Diagenode				Diagenode				
Starting amount of chromatin	1 ng	500 pg	100 pg	5 ng	1 ng	500 pg	100 pg	5 ng	

Table 1: Overview of experiment design

Table 2									
	651	**652**	**653**	**654**	**661**	**662**	**663**	**664**	**REF**
Total peaks	15052	15131	12650	15934	15747	15404	12893	16179	14946
FRIP %	75,76%	76,12%	71,27%	76,42%	75,87%	76,23%	69,57%	75,30%	65,52%
Average peak probability score	331,78	245,67	93,94	535,15	386,08	288,05	80,37	627,48	668,29
Average peak width	1070,62	1011,73	969,49	1093,42	1058,90	1021,43	895,91	1127,50	1159,79
NSC	1,369447	1,373374	1,386454	1,359902	1,366421	1,375708	1,381357	1,3476	1,284085
RSC	1,123935	1,101164	1,031533	1,164235	1,148423	1,124377	1,043203	1,206393	1,160996
Strand cross-correlation quality	high	high	high	high	high	high	high	high	high

Table 2: Quality indicators determined independently for each sample. The bottom 3 rows contain the strand cross correlation analysis, the quality is predicted by the software on a "very low - low - medium - high - very high" scale

Table 3								
	651	**652**	**653**	**654**	**661**	**662**	**663**	**664**
Overlap with Ref	100,00%	100,00%	98,02%	100,00%	100,00%	100,00%	98,25%	100,00%
Ref overlap with N	99,98%	99,98%	99,75%	99,98%	99,98%	99,98%	99,67%	99,98%
Overlap with replicate	100,00%	100,00%	99,94%	100,00%	100,00%	100,00%	99,86%	100,00%
Overlap with Illumina	99,98%	100,00%	98,14%	N/A	100,00%	100,00%	98,27%	N/A
Illumina overlap with N	100,00%	100,00%	99,95%	N/A	100,00%	100,00%	99,88%	N/A

Table 3: Overlap percentages of the top 40% of the peaks based on the ENCODE/modENCODE criteria, which requires at least 80% overlap between true replicates, in both directions

Figure 1: Excellent sequencing quality were achieved for all samples, here you can see the per base and per sequence quality scores distributions for samples 652 and 661 as examples

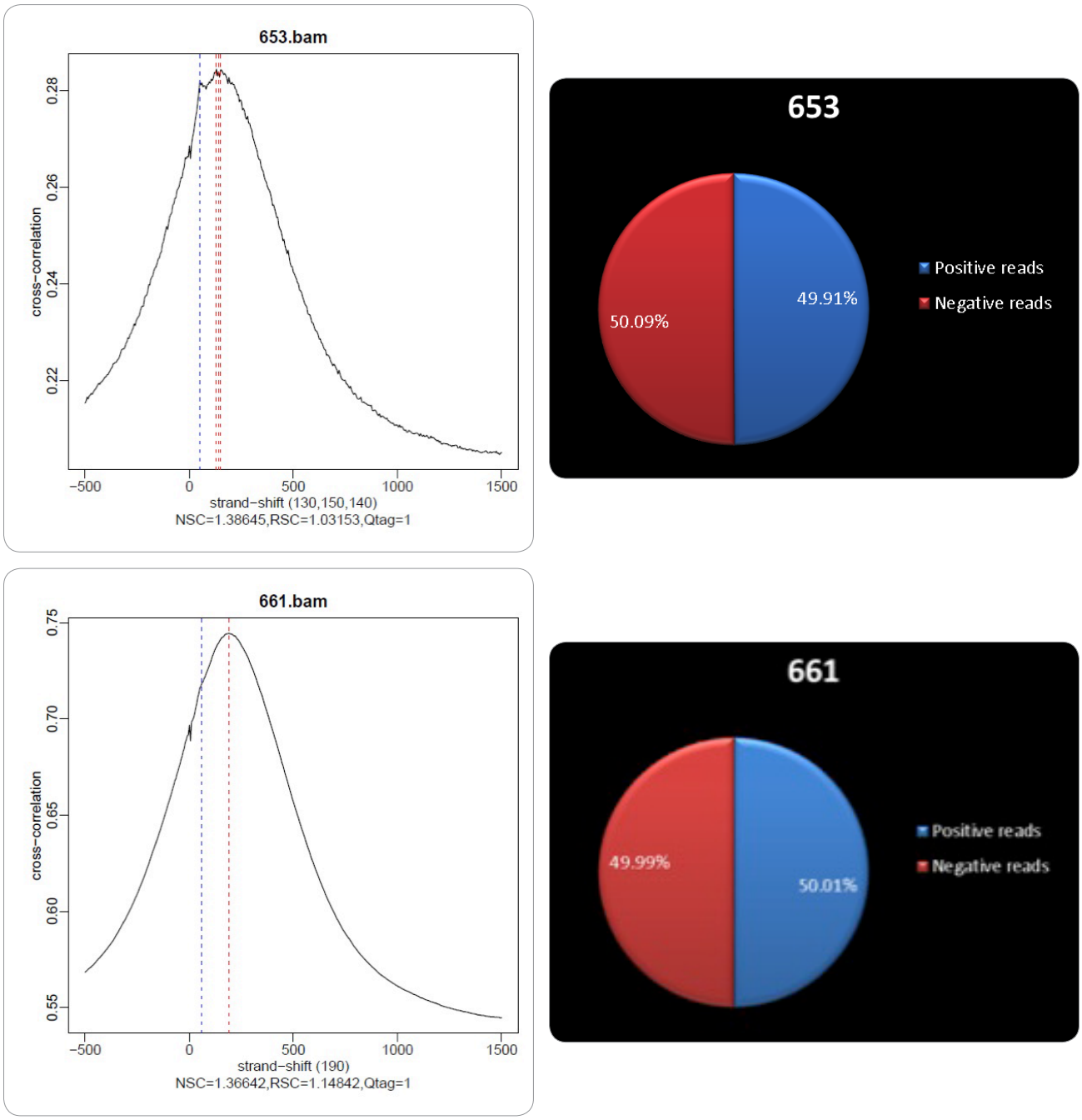

Figure 2: Strand cross correlation analyses of the datasets also showed high quality. Here are the graphs for 653 and 661 as examples; Qtag=1 means the quality is determined as "high". The pie charts show the equal distribution of sequenced reads from the positive and the negative strands

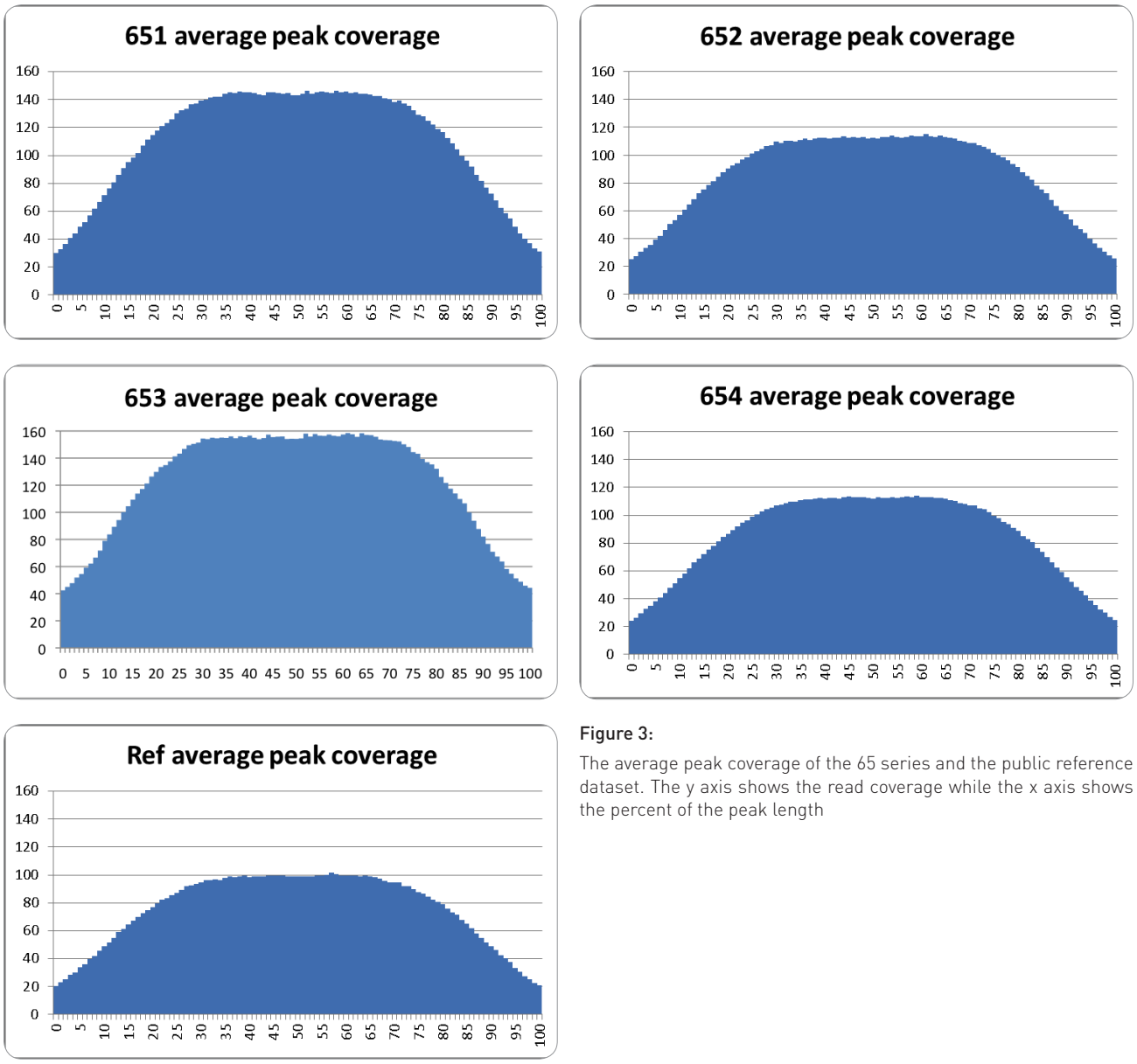

Figure 3:

The average peak coverage of the 65 series and the public reference dataset. The y axis shows the read coverage while the x axis shows the percent of the peak length

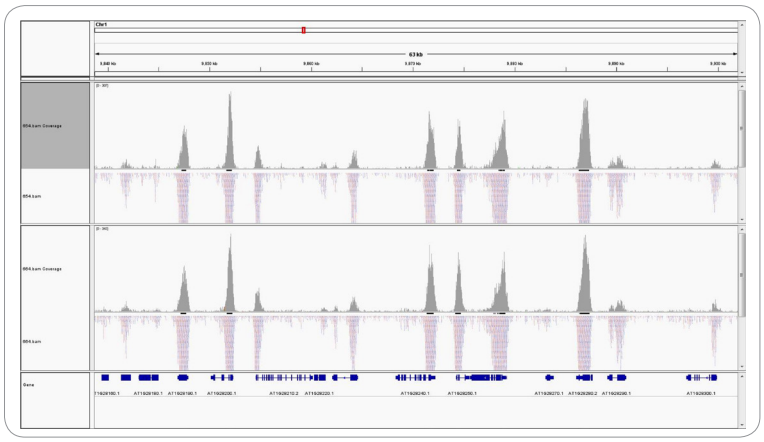

Figure 4:

Screenshot of perfectly overlapping peaks of corresponding replicate samples displayed in a viewer

Figure 5:

Average overlap profile of peaks show no bias and a central overlap pattern. The first series are the overlap profile of replicate series 2 peaks and the public reference peaks; the second series show the overlap profiles between peaks of corresponding replicate datasets

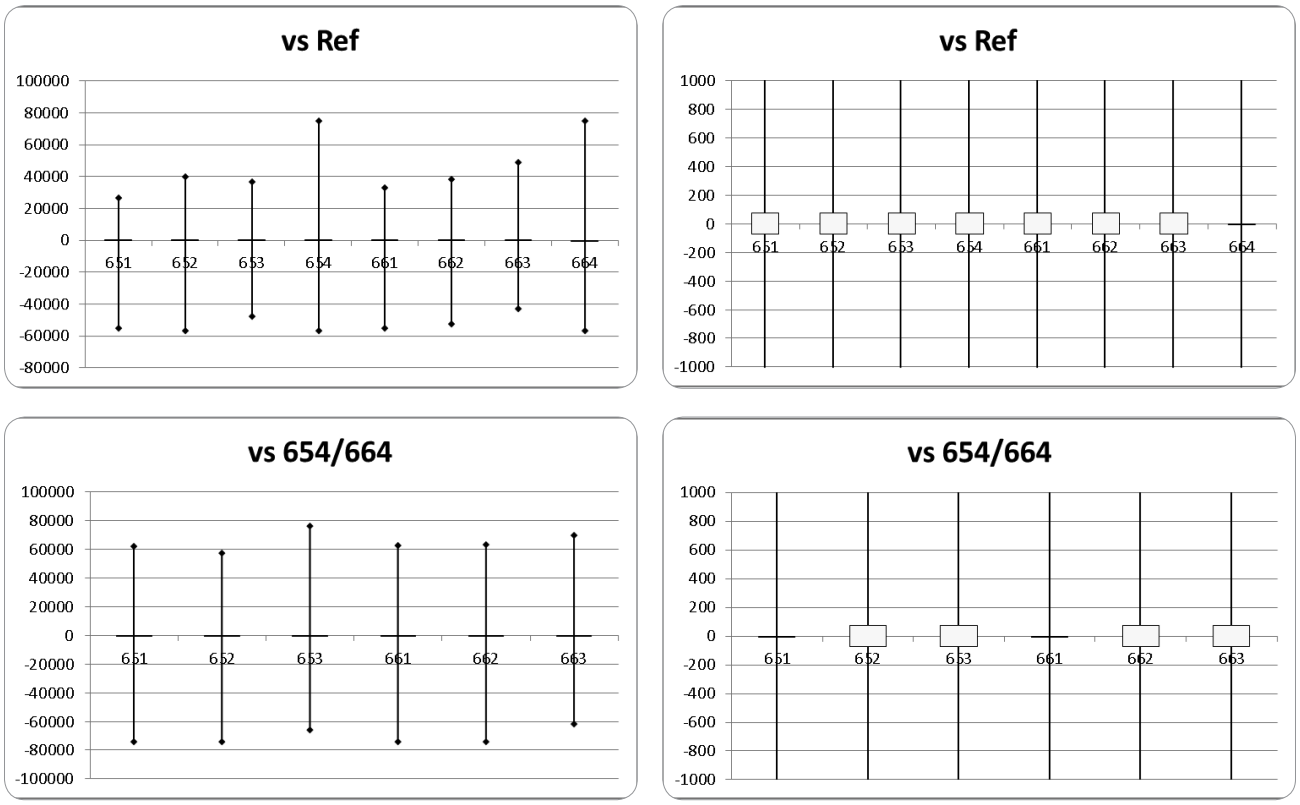

Figure 6:

Box-and-whiskers plots showing the bias free distribution of peak shifts: the significant population concentrates tightly around 0. The two charts display the shifts in the sample peaks per public reference peaks and the Diagenode peaks per Illumina peaks contexts. Zoomed in scales are presented for a better visibility of the significant population