International Journal of Computational Biology
ISSN: 2229-6700 & E-ISSN:2229-6719, Volume 4, Issue 1, 2013, pp.-56-60.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000220

COMPARING SNPS IDENTIFICATION BY CLC AND SEQMAN FROM TRANSCRIPTOME SEQUENCING DATA

SAJNANI M.R., BHATT V.D. AND JOSHI C.G.*

Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand- 388001 Gujarat, India.
*Corresponding Author: Email- bhatt_vbhv@yahoo.co.in

Received: May 17, 2013; Accepted: June 14, 2013

Abstract- Next generation sequencing (NGS) technologies produces very large amount of data at low cost. Single Nucleotide Polymorphisms (SNPs) are the most abundant form of genetic variation and there is a need to identify SNPs with less time consuming, user-friendly and accurate tools so as to minimize the efforts of researchers in analyzing data. There are various tools available for variant discovery produced from NGS data but no study presents the comparison of the SNPs discovery with CLC and SeqMan. Here we present the performance of CLC and SeqMan for prediction of potential SNPs from human buccal cancer and healthy transcriptome data obtained from Roche 454 sequencing technology based on the software utility, time, memory, disk space and accuracy of results. It was found that, performance of SeqMan seems better than that of CLC in terms of utility and accuracy. Though SeqMan required more time, memory as well as disk space than that of CLC. Both the tools are equipped with user friendly options and provide proper guideline for running.

Keywords- Next generation sequencing, Single Nucleotide Polymorphisms, Buccal, Transcriptome

Citation: Sajnani M.R., Bhatt V.D. and Joshi C.G. (2013) Comparing SNPs Identification by CLC and SeqMan from Transcriptome Sequencing Data. International Journal of Computational Biology, ISSN: 2229-6700 & E-ISSN:2229-6719, Volume 4, Issue 1, pp.-56-60.

Introduction

Single nucleotide polymorphisms (SNPs) are single base difference between haplotypes which are extensively used as genetic markers in population and conservation genetics, and molecular ecology studies [1]. Interestingly, SNPs are the most abundant form of genetic variation [2] present at regular intervals in the genome [3] and SNPs seem to be highly suitable for multiplexed genotyping assays on mass spectrometry, microarray or beadarray-based platforms [4]. Model plant species such as *Arabidopsis thaliana* (http://walnut.usc.edu/2010), *Oryza sativa* (http://irfgc.irri.org), and *Zea mays* (http://www.panzea.org/), demonstrates the potential of SNPs for extensive genome analysis which are suitable for genome wide association studies and molecular breeding concepts like genomic selection [4]. Currently, there are several approaches available for calling SNPs from NGS data, including CLC Genomics workbench (CLC Bio, Aarhus, Denmark), SeqMan (DNASTAR Inc., Madison, WI), Pyrobayes [5,6], PolyBayes, SOAPsnp [7], Varscan [8], SNVMix [9,10], SeqEM [11], MAQ [12] and Atlas-SNP2 [10]. Pyrobayes and PolyBayes recalibrate base calling from raw data, and then implement a Bayesian approach that incorporates prior information with population mutation rates to detect SNP. MAQ derives using Bayesian statistical model measures the confidence that a read actually comes from the position it aligns to, error probabilities from the raw sequence quality scores, sampling of the two haplotypes, and an empirical model for correlated errors at a site. SOAPsnp and SeqMan are also based on the Bayes' theorem. It first recalibrates the sequencing quality score to calculate the likelihood of genotype for each position with existing conversion matrix, and then combines the prior probability for each genotype to infer the true genotype. Varscan uses parameters such as the overall coverage, the number of supporting reads, average base quality, and the number of strands observed for each allele to predict genotypes [7,9]. SNVMix combines three Binomial-mixture models to model allelic counts, nucleotide and mapping qualities of the reads and infers SNPs and model parameters with the expectation maximization (EM) algorithm. In contrast, SeqEM uses the EM algorithm to numerically maximize the observed data likelihood with respect to genotype frequencies and the nucleotide-read error rate based on the NGS data of multiple unrelated individuals [9,11]. Atlas-SNP infers systematic errors of base substitutions on single reads by fitting training datasets using a logistic regression model which identified read sequence-related covariates to the base-quality score [10]. CLC calls SNPs using Neighborhood Quality Standard (NQS) [13] where sequence quality of the varying base is based on the quality of the neighborhood bases [14]. No study provides a comparison of CLC and SeqMan for SNPs detection, and so in the present study we aim to find SNP from buccal cancerous and healthy tissue transcriptome data obtained by Roche 454 pyrosequencing to represent the performance of CLC and SeqMan softwares thereby presenting how different algorithms treat individual variations.

Methods

NGS reads of buccal cancerous and healthy tissue transcriptome were obtained from Roche 454 pyrosequencing technology and processed in GS Run Processor v.2.5 for base calling on linux operating system (OS). The reads were transferred to windows system for quality filtering, mapping and SNPs identification.

System Configuration

The pipeline for SNP detection was carried out on Windows 7 edition consisting of 16.0 GB installed memory, Intel® Xeon® CPU X 5650 @ 2.66 GHz (24 system processors), 2 TB hard disk and 64 bit OS.

Quality Filtering

Quality filtering of the reads obtained from, both cancer and healthy tissues were performed using NGSQC Toolkit [15] with default parameters described in [Table-1].

Table 1- Quality trimming parameters

Primer/Adaptor library	Rapid Library (Standard)
Homopolymer trimming	On
Length of the homopolymer to be removed	8
Length filter	On
Cut-off for minimum read length	40
Cut-off read length for HQ	70%
Cut-off quality score	20
Only statistics	Off
Number of processes	1

Reference Mapping

Reference genome of *Homo sapiens* assembly GRCh37.p5/hg19 and dbSNP build 135 was downloaded from DNASTAR genome package (http://www.dnastar.com/t-dbsnp_files.aspx) for mapping and SNPs detection. To carry out mapping we used ‘map read to reference’ application of CLC and ‘templated assembly’ application of SeqMan for aligning buccal reads of cancerous and healthy transcriptome to reference genome individually.

SNP Detection

To run SNP detection program with CLC, default parameters were assigned. CLC works with NQS algorithm. Thus window size assigned was 11, number of gaps and mismatches within window length of the read was 2, average quality score of the nucleotides in a read within the specified window length was 15 for the base to be included in the SNP, and quality score for the central base was 20. To avoid SNPs calling in areas of low coverage, where one would get a higher amount of false positives, minimum number of valid reads at particular position was taken as 4 whereas 35% considered as minimum variant frequency to validate reads at this position with different base. To compare of the performance of CLC Genomic Workbench we performed mapping and SNP detection with 'templated assembly' application using SeqMan v.4.0.0. It uses Bayesian statistical model to call SNPs. Parameters were adjusted according to the CLC parameters where mapping similarity, maximum number of gaps, minimum coverage, minimum variant frequency and minimum SNP threshold referred as quality of central base in SeqMan was kept similar to CLC parameters except minimum base quality score which was kept 10. Mapping and SNPs parameters are described in [Table-2]. True positives (TP) and false positive (FP) rates of SNP was calculated with following formula:

$$TP = (TP/\text{total SNPs}*100) \text{ and } FP = (FP/\text{total SNPs}*100)$$

Where:
TP: number of SNPs matched to dbSNP
FP: number of SNPs did not matched to dbSNP
Total SNPs: number of SNPs identified by CLC or SeqMan

Table 2- Mapping and SNPs parameter used for analysis

Mapping Parameter	CLC	SeqMan
Min. mer size	-	21
Length fraction (min aligned length)	50%	50
Similarity	80%	80%
Mismatch cost	2	20
Max no of gaps	2	20
gap penalty	-	20
Insertion cost	3	-
SNPs Parameter		
SNP Window length	11	-
Min coverage	4	4
Min variant frequency	35%	35%
SNP confidence threshold/min quality score of central base	20	20
Min Base quality score	15	10
Algorithm	neighbourhood quality standard	bayesian

Results

Buccal Tissue Transcriptome Data

We used a Roche 454 FLX instrument to generate ‘Titanium chemistry’ reads from cDNA libraries from two different human buccal tissue types, cancerous and healthy tissue transcriptome, yielded 31.5 and 37.9 million base pairs (Mb) having 106113 and 130148 sequence reads with an average read length of 297 and 291 nucleotides, respectively [16].

System Requirements

To identify variants with different software like CLC and SeqMan we performed analysis on windows OS. We found that to perform analysis on CLC and SeqM minimum 256 megabyte (MB) and 16 gigabyte (GB) Random Access Memory (RAM) is required respectively. CLC require Intel or AMD CPU whereas SeqMan require Quad-Core 2 GHz. Both the software requires 64 bit processor and windows operating system. Our pipeline for SNPs identification was carried out on windows operating system having Windows 7 edition, 16 GB RAM and 2 terabyte (TB) hard disk. [Table-3] describes the features of the software used in the present study.

Quality Filtering

We performed quality filtering and homopolymer trimming on NQC Toolkit and found 93589 cancerous and 114509 healthy reads with 22734524 bp and 26605464 bp having 30 as an average base quality score. Thus total of 49.33 MB of data from cancerous and healthy tissue transcriptome was taken for mapping and SNP analysis. Approximately 88.45% reads of cancerous and 87.98% reads of healthy tissue were found with high quality having 367 and 360 N50 read size, respectively presented in [Table-4] and [Table-5].

Comparison of SNPs Discovery with CLC and SeqMan

Reference Mapping

For SNP analysis, CLC parameters were kept default and SeqMan parameters were set accordingly. Firstly, we carried out mapping of high quality transcriptome reads of buccal cancerous and healthy tissue with reference genome package of *Homo sapiens* GRCh37.p5/hg19 using CLC and SeqMan. Mapping with reference, cancer tissue reads showed 4.02% (124446232 bp) and 9.74%

(301507400 bp) human genome coverage whereas healthy tissue reads showed 3.01% (93179890 bp) and 7.64% (236586197 bp) human genome coverage using CLC and SeqMan, respectively. We observed that while mapping in CLC, cancerous and healthy reads showed highest similarity with chromosome 17 (0.32%) and chromosome 17 (0.23%) whereas while mapping in SeqMan, cancerous and healthy reads showed highest similarity with chromosome 4 (1.13%) and chromosome Y (0.92%), respectively.

Table 3- Features of SNP program compared in this study

Tool	Application	Algorithm	Author	URL
CLC Genomics Workbench 4.9.0	SNP Detection	Neighbourhood quality standard	CLC	http://www.clcbio.com/
DNA STAR 4.0.0	SeqManNgen	Bayesian	SeqMan	http://www.dnastar.com/

Table 4- Qualtity statistics of NGSQC Toolkit on buccal cancerous and healthy tissue transcriptome data

Quality statistics	Cancerous	Healthy
Total number of reads	106113	130148
Total number of trimmed reads containing homopolymer	11188	16235
Total number of trashed reads (length <40 bp after trimming)	1037	1513
Total number of low quality reads (excluding <40 reads)	11217	14126
Total number of HQ reads	93888	114509
Percentage of HQ reads	88.48%	87.98%
Total number of bases	31668249	38086452
Total number of bases in HQ reads	27427490	32413385
Total number of HQ bases in HQ reads	22736614	26605953
Percentage of HQ bases in HQ reads	82.90%	82.08%
Number of Primer/Adaptor trimmed reads	77	15
Total number of HQ filtered reads	93859	114509
Percentage of HQ filtered reads	88.45%	87.98%

Table 5- Results of quality trimming before and after filtering buccal transcriptome data

File name	Cancer Raw	Cancer filtered	Normal Raw	Normal filtered
Total number of reads	106113	93859	130148	114509
Minimum read length	40	40	40	40
Maximum read length	717	605	725	634
Average read length	298.44	292.19	292.64	283.06
Median read length	307	297	294	278
N25 length	447	440	444	436
N50 length	374	367	369	360
N75 length	283	272	268	253
N90 length	197	189	184	177
N95 length	144	142	144	138
Total number of bases	31668249	27425000	38086452	32412803
Total number of HQ* bases	25404734	22734524	30134755	26605464
Percentage of HQ* bases	80.22%	82.90%	79.12%	82.08%
Average quality score	29.02	30	28.71	29.72

**HQ: High quality*

SNP Analysis

We found total 3153 and 491 SNPs in cancerous tissue reads whereas 2529 and 471 SNPs in healthy tissue read by CLC and SeqMan, respectively. Using PERL (Practical Extraction Report Language) scripts we calculated common SNPs detected between CLC and SeqMan for cancerous and healthy reads separately. Out of 3153 and 491 SNPs, 326 SNPs were common in cancerous tissue reads whereas out of 2529 and 471 SNPs, 116 SNPs were common in healthy tissue reads. The 12.4% and 56% SNPs identified in cancerous tissue and 15.93% and 66.48% SNPs identified in healthy tissue by CLC and SeqMan, respectively were common with dbSNP. Thus these common SNPs were considered as true positives whereas all uncommon SNPs were considered as false positives. [Table-6] represents the detailed SNP results on buccal transcriptome data.

Table 6- SNP detection results from CLC and SeqMan on buccal cancerous and healthy transcriptome data

Tissue type	Cancerous tissue		Healthy tissue	
Software type	CLC	SeqMan	CLC	SeqMan
Total bases mapped	124446232	301507400	93179890	236589197
Total bases mapped (%)	4.02%	9.74%	3.01%	7.64%
Total SNPs	3153	491	2529	471
Common with dbSNPs	12.40%	56%	15.93	66.48
True Positives	391	275	403	299
Percent True Positives	12.4	56.01	15.93	66.48
False Positives	2762	216	2126	172
Percent False Positives	87.59	43.99	84.06	36.51
Common SNPs	326		116	

Time Calculation and Memory Usage

Time taken for mapping and SNP analysis was calculated for both the softwares. CLC utilized approximately 27 minutes whereas SeqMan utilized approximately 2 hours 15 minutes. It was observed that the mapping required more time than SNP identification. CLC utilized approximately 26 minutes for mapping which aligned 4.02% and 3.01% reads in cancerous and healthy respectively. Compared to CLC, SeqMan utilized 43% more time to perform alignment but with 41.27% more reads aligned in 1 hours and 45 minutes. Total disk space consumed by CLC was 0.98 GB whereas SeqMan consumed 138 GB for mapping and SNP analysis.

Discussion

SNPs have a wide variety of applications in biological research. To obtain best set of variants, there is need to use right combination of tools to discover them with less false positive calls obtained due to amplification bias and sequencing error. Efficient SNP discovery and genotyping in a highly heterozygous genome containing a high proportion of repetitive elements and paralogous sequences is difficult [17]. In order to make NGS technology ubiquitous and clinically useful, one needs to come up with simplified analysis tools that produce more true positive calls and reduces efforts and money required for downstream validation experiments [18]. This will allow biologists focus more on their work rather than on optimizing analytical tools for variant discovery. Here, we present a comparative study of two different commercially available applications for read alignment and variant discovery obtained by CLC and SeqMan which applies NQS and Diploid Bayesian algorithms, respectively for variant discovery.

Quality Trimming

The quality of data is very important for various downstream analyses such as sequence assembly, mapping, single nucleotide polymorphisms identification and gene expression studies. Thus availability of accurate base quality data could improve the accuracy of SNP detection. Moreover homopolymers are the major problem observed in Roche 454 sequencing. The signal intensity distribution broadens with the length of the homopolymer which leads to an ambiguous base call which may further lead to frame-shift affecting the downstream processing [15]. Thus we carried out quality check and homopolymer removal on the sequence data for which we used NGSQC Toolkit to provide high quality data which makes reads suitable for SNP detection giving less false positives SNP count with more accuracy. We observed approximately similar percent of reads 88.45% of cancerous and 87.98% of healthy tissue transcriptome, passing the quality check and homopolymer trimming for downstream analysis. It is reported that that the average quality score of a read is inversely proportional to the number of errors in that read [19]. Our read showed average 30 quality score having fewer error rates.

Alignment to Reference Sequences

We took high quality reads for mapping with latest reference of *Homo sapiens*. For both the software stringent parameters were used for mapping and SNPs detection analysis to reduce the false positive rates. Firstly we observed that CLC do not provide reference package whereas SeqMan has systematic reference package for mapping and for SNP identification. Moreover in our results, we observed percentage of reference covered during mapping was approximately 39-41% higher with CLC than that of SeqMan for cancerous and healthy reads. CLC genomics keep percent of minimum aligned length to 50% whereas SeqMan specifies the alignment length to 50 bases. In this case the possibility of alignment is reduced with increase in read length whereas in SeqMan increase in read length do not reduce the alignment as no matter what is the size of the read but it will align minimum 50 bases which thereby increase the chances of alignment.

SNP Analysis

CLC and SeqMan carry out SNP detection with two different well known approaches NQS and Bayesian model, respectively. In NQS [20], SNP detection looks at each position in the mapping to determine if there is a SNP at a particular position. In order to make a qualified assessment, it also considers the general quality of the neighboring bases by keeping 11 window size to determine that 5 bases from left and right should have minimum quality score 15 and the central base should have minimum quality score 20 to determine it as candidate SNPs. It was hypothesized that bases surrounded by perfectly aligned, consistently high-quality sequence (termed 'good neighborhoods') might be more accurate than predicted by PHRED [21]. Thus NQS was used to identify such bases [22]. SeqMan applies Bayesian algorithm to call SNPs between homozygous reference, homozygous variant and heterozygous. This algorithm first run simple SNP caller on each column of the read. If the column passes a minimum percentage screen, it then check against a minimum variant depth where the most frequent variant base must meet or exceed this threshold. Thereafter putative SNP containing columns are evaluated with a statistical model that considers the two most frequent bases in the column as possible alleles. The model then calculates the P value of each set of bases which will be based on the base frequency, combined frequency of the two bases, the quality scores and the directions of the reads. Moreover the heterozygous call's probability is based on simple permutations and a constant modifier, with the strands considered separately. Since they are the only possible genotypes, probabilities are normalized against one another, and the highest probability is called [23].

In present study, we observed the percentage of SNPs in SeqMan reduced to 84.42% and 81.37% compared to CLC in cancerous and healthy tissue transcriptome, respectively. We also observed that the number of SNPs that matched to dbSNPs were higher in SeqMan than CLC. Incorrectly detected SNPs are primarily due to paralogous gene sequences interfering with the assembly of short NGS reads [17]. It is reported that dbSNP genotypes for prior probability calculation helps in distinguishing real heterozygotes from errors in regions of low-depth sequencing. The use of additional information for prior probability under the general Bayesian probability framework could likely aid in further improving accuracy of posterior probability calculation [7]. The performance of NQS depends on the read quality in NQS windows with low coverage data. NQS for SNP detection is appropriate for any sequencing system that has suitable quality scores whereas Bayesian models depends on P value which is based on the base frequency, combined frequency of the two bases, the quality scores and the directions of the reads. Bayesian algorithm combines a priori knowledge about the sequence context with the specific, observed data represented by the sequences under examination. Such prior knowledge includes an approximate average polymorphism rate in the region, and the expected ratio between transitions and transversions thereby decreasing rate of false positive polymorphisms [1].Thus it seems that the stringency for SNP detection is higher in Bayesian model than NQS which ultimately decreases the number of SNPs in SeqMan. We observed 326 and 116 SNPs common in CLC and SeqMan in cancerous and healthy tissue respectively. The use of these strict parameters should allow reducing the false positive rate and assuring relatively high quality results. In terms of utility we found that SeqMan can provide best results which includes prediction of SNPs in coding and noncoding region, synonymous and non-synonymous SNPs, SNPs that match dbSNP, SNPs that affect frame shift mutations as well as it provide post analysis application with ArrayStar whereas CLC only predicts amino acid change. SeqMan SNP analysis can be performed on Windows OS but on CLC, it can be carried out on Linux, Mac, and Windows OS. Both the tools are equipped with user friendly options and provide proper guideline for running. Though SeqMan require large disk space and it was observed to be more time consuming than CLC but it gives accurate results with match to db SNP. Lastly proportion of SNP discovery in such a high-throughput screen may not represent true polymorphisms [1]. Further analyses are needed to validate these SNPs before their use in population structure analyses. Methods for validation of the SNPs will vary with the specific aims of researchers and may make use of array based technologies (Affymetrics, Illumina), or platforms that use allele specific primer extension/ligation based methods as well as resequencing by the Sanger method for smaller scale approaches [24].

Conclusion

In brief, the overall performance of SeqMan seems better than that of CLC in terms of utility and accuracy. Though SeqMan required more time, memory as well as disk space than that of CLC but runs on widely used Bayesian model considering dbSNP for SNP analysis to avoid false positives and is suitable for high coverage data, whereas, CLC runs on NQS for SNP detection which is appropriate for any sequencing system that has high quality scores. Both the tools are equipped with user friendly options and provide proper guideline for running.

Acknowledgments

Authors would like to acknowledge Dr. A.K. Tripathi, Dr. A.K. Patel and Dr. M.R. Upadhyay for suggestions in analysis.

References

[1] Kuhl H., Tine M., Hecht J., Knaust F., Reinhardt R. (2011) *Comp. Biochem. Physiol. Part D Genomics Proteomics,* 6, 70-5.

[2] Rafalski A. (2002) *Curr. Opin. Plant Biol.*, 5, 94-100.

[3] Ponting R.C., Drayton M.C., Cogan N.O., Dobrowolski M.P., Spangenberg G.C., Smith K.F., Forster J.W. (2007) *Mol. Genet. Genomics*, 278, 585-97.

[4] Gupta P.K., Rustgi S., Mir R.R. (2008) *Heredity (Edinb)*, 101, 5-18.

[5] Cock P.J., Fields C.J., Goto N., Heuer M.L., Rice P.M. (2010) *Nucleic Acids Res.*, 38, 1767-71.

[6] Quinlan A.R., Stewart D.A., Stromberg M.P., Marth G.T. (2008) *Nat. Methods,* 5, 179-81.

[7] Li R., Li Y., Fang X., Yang H., Wang J., Kristiansen K. (2009) *Genome Res.*, 19, 1124-32.

[8] Koboldt D.C., Chen K., Wylie T., Larson D.E., McLellan M.D., Mardis E.R., Weinstock G.M., Wilson R.K., Ding L. (2009) *Bioinformatics*, 25, 2283-5.

[9] Goya R., Sun M.G., Morin R.D., Leung G., Ha G., Wiegand K.C., Senz J., Crisan A., Marra M.A., Hirst M. (2010) *Bioinformatics*, 26, 730-6.

[10] Shen Y., Wan Z., Coarfa C., Drabek R., Chen L., Ostrowski E.A., Liu Y., Weinstock G.M., Wheeler D.A., Gibbs R.A. (2010) *Genome Res.,* 20, 273-80.

[11] Martin E.R., Kinnamon D.D., Schmidt M.A., Powell E.H., Zuchner S., Morris R.W. (2010) *Bioinformatics*, 26, 2803-10.

[12] Li H., Ruan J., Durbin R. (2008) *Genome Res.*, 18, 1851-8.

[13] Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., Lander E.S. (2000) *Nature*, 407, 513-6.

[14] Matukumalli L.K., Grefenstette J.J., Hyten D.L., Choi I.Y., Cregan P.B., Van Tassell C.P. (2006) *BMC Bioinformatics*, 7, 4.

[15] Patel R.K., Jain M. (2012) *PLoS One*, 7, e30619.

[16] Sajnani M.R., Patel A.K., Bhatt V.D., Tripathi A.K., Ahir V.B., Shankar V., Shah S., Shah T.M., Koringa P.G., Jakhesara S.J., Koringa P.G., Joshi C.G. (2012) *Gene*, 507, 152-8.

[17] Studer B., Byrne S., Nielsen R.O., Panitz F., Bendixen C., Islam M.S., Pfeifer M., Lubberstedt T., Asp T. (2012) *BMC Genomics*, 13, 140.

[18] Pattnaik S., Vaidyanathan S., Pooja D.G., Deepak S., Panda B. (2012) *PLoS One*, 7, e30080.

[19] Huse S.M., Huber J.A., Morrison H.G., Sogin M.L., Welch D.M. (2007) *Genome Biol*, 8, R143.

[20] Unneberg P., Stromberg M., Sterky F. (2005) *Bioinformatics*, 21, 2528-30.

[21] Ewing B., Hillier L., Wendl M.C., Green P. (1998) *Genome Res.*, 8, 175-85.

[22] Brockman W., Alvarez P., Young S., Garber M., Giannoukos G., Lee W.L., Russ C., Lander E.S., Nusbaum C., Jaffe D.B. (2008) *Genome Res.*, 18, 763-70.

*[23]*Shah S.P., Morin R.D., Khattra J., Prentice L., Pugh T., Burleigh A., Delaney A., Gelmon K., Guliany R., Senz J. (2009) *Nature*, 461, 809-13.

[24] Ding C., Jin S. (2009) *Methods Mol. Biol.,* 578, 245-54.