Published online 1 March 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 12* ***3999–4010***
doi:10.1093/nar/gkq116

Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing

Fabian Birzele[1,*], Jochen Schaub[2,*], Werner Rust[1], Christoph Clemens[2], Patrick Baum[1], Hitto Kaufmann[2], Andreas Weith[1], Torsten W. Schulz[2] and Tobias Hildebrandt[1]

[1]Department of Pulmonary Research, Group Genomics and [2]Department of Biopharmaceutical Process Science, Boehringer Ingelheim Pharma GmbH & Co KG, Birkendorferstraße 67, 88397 Biberach an der Riß, Germany

Received December 23, 2009; Revised February 5, 2010; Accepted February 6, 2010

ABSTRACT

The arrival of next-generation sequencing (NGS) technologies has led to novel opportunities for expression profiling and genome analysis by utilizing vast amounts of short read sequence data. Here, we demonstrate that expression profiling in organisms lacking any genome or transcriptome sequence information is feasible by combining Illumina's mRNA-seq technology with a novel bioinformatics pipeline that integrates assembled and annotated Chinese hamster ovary (CHO) sequences with information derived from related organisms. We applied this pipeline to the analysis of CHO cells which were chosen as a model system owing to its relevance in the production of therapeutic proteins. Specifically, we analysed CHO cells undergoing butyrate treatment which is known to affect cell cycle regulation and to increase the specific productivity of recombinant proteins. By this means, we identified sequences for >13 000 CHO genes which added sequence information of ~5000 novel genes to the CHO model. More than 6000 transcript sequences are predicted to be complete, as they covered >95% of the corresponding mouse orthologs. Detailed analysis of selected biological functions such as DNA replication and cell cycle control, demonstrated the potential of NGS expression profiling in organisms without extended genome sequence to improve both data quantity and quality.

INTRODUCTION

Development of next generation sequencing (NGS) platforms such as Illumina's Genome Analyzer (Solexa Sequencing), Roche's 454 method or the ABI Solid Sequencers have provided novel tools for expression profiling and for genome analysis (1). Each technology has different properties with respect to lab handling, read length and quality, and sequence output. Also, the chosen methodology has implications on subsequent data analysis that may be a considerable challenge. Only recently, current available NGS methods have been described in detail in the reviews by Metzker (2) or Shendure *et al.* (3). The Illumina Genome Analyzer platform used in this study allows to sequence millions of (relatively short) reads in parallel, resulting in the generation of substantial amounts of mRNA or DNA sequence data in only one single experiment, and is especially well-suited to perform sensitive (very detailed) transcriptome analyses. NGS methods have already been shown to address a large variety of different problems, ranging from reliable expression profiling and splice variant analysis in organisms where reference genomes are known (4–7), the detection of sequence and structural variations in the human genome (8) and the characterization of new transcription factor binding motifs (9) to the analysis of folding principles of the human DNA in the nucleus (10).

Here we applied NGS for gene expression profiling in Chinese hamster ovary (CHO) cells. Despite the fact that CHO cells are widely used for the production of therapeutic proteins (mainly monoclonal antibodies), there is currently no comprehensive sequence information describing their genome or transcriptome. Recombinant

*To whom correspondence should be addressed. Tel: +49 (7351) 54-5283; Fax: +49 (7351) 54-5991;
Email: fabian.birzele@boehringer-ingelheim.com
Correspondence may also be addressed to Jochen Schaub. Tel: +49 (7351) 54-96550; Fax: +49 (7351) 54-98391;
Email: jochen.schaub@boehringer-ingelheim.com

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

antibodies have become highly important therapeutic agents in the last decade and their demand is rapidly increasing. They are, for example, currently used in the treatment of a variety of oncology and inflammatory diseases (11) and are usually produced in mammalian cell culture to achieve the extensive post-translational modifications such as glycosylation that is required for optimal function in terms of half-life, stability, antibody-dependent cell-mediated cytotoxicity (ADCC) and complement-dependent cytotoxicity (CDC). Given this high demand, there is a need to improve process efficiency in antibody production. Therefore, a better understanding of the biology of the production cell lines is a key factor (12,13). However, despite their importance, little is known about the complex intracellular processes in CHO cells, for example, changes in the transcriptional landscape. Such large-scale datasets would enable both a detailed analysis of a specific phenotype of a certain cell clone (e.g. cell-specific productivity) and a comprehensive molecular picture of the cellular responses to environmental changes such as a change in the composition of cell culture media (14). Thus, these data could greatly help to improve cell lines and production processes to finally obtain high recombinant product concentrations of correctly glycosylated antibodies.

The major drawback for the application of genomics approaches in Chinese hamster cell lines so far is given by the fact that the complete genome sequence is not available. This makes (powerful) large-scale expression profiling with standard microarray platforms difficult. Recently, considerable progress has been achieved by large-scale expressed sequence tag (EST) sequencing of the CHO transcriptome, which has resulted in a custom-made CHO-specific Affymetrix microarray (15,16). This array currently detects gene expression of ~10 000 CHO genes. In general, this approach suffers from two limitations. First, only a fraction of the expected number of the expressed genes in CHO cells is likely to be present on the chip, as they have not been detected by EST sequencing yet. Second, chip probe design without the complete genome sequence is difficult, as reliable genome information is mandatory to avoid cross-hybridization effects between two or more genes. For other important model organisms such as the minipig or cynomolgus, no information on the genome or transcriptome level is available, making chip design impossible.

In this study, CHO mRNA sequencing using Illumina's GAII was carried out to demonstrate the feasibility of performing reliable and detailed expression analysis of organisms without an appropriate reference genome, solely based on the information of the genomes and transcriptomes of related species (mouse and rat). Moreover, we established a computational workflow for pre-processing of the CHO NGS data that greatly supported subsequent expression analysis steps.

In particular, we propose to perform a transcriptome assembly of the NGS reads in the first step, in order to obtain longer CHO sequence contigs. Those contigs are likely to represent the true CHO sequences, and therefore capture mutations, insertions and deletions, which are present in the Chinese hamster in comparison with related species. In the second step, the contigs can be reliably annotated to a reference genome (in our case the mouse genome) to assign the respective orthologous genes and to annotate potential functions. Read analysis is performed as the third step by combining mapping information from mouse, rat and CHO assembly sequences to obtain final read counts for CHO transcripts and genes.

As an exemplary study, we show that our workflow allows high-throughput and large-scale expression profiling of CHO gene expression. To this end, we investigated the effect of sodium butyrate treatment, since it is relevant for biotechnology applications and cell biology. Sodium butyrate is an important supplement in mammalian cell culture to increase the specific productivity of recombinant proteins in CHO cells (17). It has also been analysed in the context of oncology as an inhibitor of cell cycle progression and as an inducer of apoptosis in cancer cell lines (18,19).

Sodium butyrate inhibits histone deacetylases leading to a subsequent increase in the accessibility of the DNA to transcription factors. Several studies have already analysed the effects of sodium butyrate treatment on different cell lines (17–21) and found that, among other effects, sodium butyrate mediates a down-regulation of cell cycle proteins followed by an arrest of the cells in the G1 or G2 phase (22).

Our analysis revealed two major advantages of applying NGS for CHO expression profiling. First, biologically meaningful expression analysis of CHO cells is possible using NGS data. Many of the cellular processes and genes leading to sodium butyrate-mediated cell cycle arrest could be identified in a much greater detail compared with a chip platform. Genome-wide expression profiling by NGS can be performed without the time and cost-intensive steps to compile a set of EST sequences, and the error-prone design of custom-made expression arrays in the absence of the complete genome sequence information. Second, NGS can provide a significant amount of novel sequence information on CHO transcripts, which can be used for further NGS studies or to gain a deeper understanding of the CHO genome and transcriptome. Sequencing data from 12 samples allowed for the assembly of more than 6000 CHO transcripts that were likely to be complete with respect to their orthologs in mouse. Moreover, gene expression of more than 13 000 genes could be profiled.

Finally, this study demonstrates the potential of a novel bioinformatics pipeline combined with NGS data for the analysis of other model organisms where no reference genomes are available, but for which large-scale expression profiling would reveal an abundance of novel information. These data will be required to globally elucidate complex molecular networks in those genomically 'unknown' organisms.

MATERIALS AND METHODS

Cell culture, butyrate treatment and analytical methods

A recombinant IgG producing CHO cell line was cultivated in a controlled fed batch process. The bioreactor

(5.5 l starting volume, 37°C cultivation temperature) was inoculated with a viable cell concentration of 3.0×10^5 cells/ml. The pH was controlled at 7.1 and the dissolved oxygen concentration at 60% air saturation by adjustment of stirrer speed and aeration. Proprietary, chemically defined, serum-free basal and fed batch media were used. Three cultivations using the same inoculum pre-culture were performed in parallel. No butyrate was added to the control culture, in the other two cultivations butyrate was added at concentrations of 0.5 mM, respectively, 1.0 mM, at cultivation Day 5.25.

Cell concentration and cell viability were determined by the trypan blue exclusion method with a CEDEX cell analyzer (Innovatis AG, Bielefeld, Germany). Recombinant IgG antibody concentration was quantified by surface plasmone resonance detection with a Biacore C instrument (GE Healthcare Europe GmbH, Munich, Germany). Samples for gene expression analysis were taken at cultivation Days 0, 4, 6 and 8.

RNA isolation

RNA isolation was carried out using a MagMAX Express-96 Magnetic Particle Processor (Ambion, Austin, TX, USA) and the MagMAX-96 Total RNA Isolation Kit (Ambion, Austin, TX, USA) according to the manufacturer's protocol. Total RNA concentration was quantified by Nanodrop (NanoDrop Technologies, Wilmington, DE, USA). RNA quality was characterized by the quotient of the 28S to 18S ribosomal RNA electropherogram peak using an Agilent 2100 bioanalyzer and the RNA Nano Chip (Agilent Technologies, Santa Clara, CA, USA).

Library preparation and sequencing

All libraries were prepared using the mRNA-Seq 8 sample prep Kit (Illumina, San Diego, CA USA) according to the manufacture's instruction. In brief, first, magnetic beads containing poly-T molecules were used to purify mRNA from 5 µg of total RNA. Second, samples were chemically fragmented and reverse transcribed into cDNA. Finally, end repair and A-base tailing was performed before Illumina adapters were ligated to the cDNA fragments. After a gel size fractionation step to extract fragments of ~200 bp, 30 µl of the purified samples were amplified by 15-cycle PCR. Amplified material was validated and quantified using an Agilent 2100 bioanalyzer and the DNA 1000 Nano Chip Kit (Agilent, Technologies, Santa Clara, CA, USA).

Libraries were loaded onto the channels of the flow cell at 5–7 pM concentration. Sequencing was carried out on the Genome Analyzer II (Illumina, San Diego, CA, USA) by running 36 cycles using Illumina's Single Read Cluster Generation Kit and 36 Cycle Sequencing Kit according to the manufacturer's instructions.

Read mapping

Short read sequences were mapped to mouse and rat transcripts obtained from Ensembl release 55 (23) [processed by the ProSAS pipeline (24)] using the Bowtie mapping algorithm (25). A maximum of four mapping errors were allowed to account for the larger number of mutations to be expected between CHO reads and the two reference transcriptomes.

Mapping to assembled CHO contigs was also performed with stricter mapping criteria of at most two mismatches between CHO contig and a read. It is important to note that Bowtie does not allow for insertions and deletions to occur in the alignment between reference sequences and read such that all matches are gapless.

CHO transcriptome assembly

Assembly strategies. To obtain longer CHO mRNA sequences, which are useful in subsequent analysis steps, two different assembly strategies were applied and combined in a final CHO assembly. First, we computed two *de novo* assemblies of all reads pooled for each of the two flow cells using Velvet (26). This led to an assembly of the read data which is not limited to and biased towards sequences known in a reference genome like in mouse or rat, and may also contain contigs which are unique for CHO, like poorly conserved transcript UTRs or novel genes.

The second assembly strategy, which will be called knowledge-based assembly, makes use of all known Ensembl mouse transcripts (or the corresponding rat orthologs) and all reads which have been mapped to those sequences. Knowledge-based assembly is performed by collecting all reads mapping to a specific mouse gene in any of the 12 lanes and running Velvet on those short reads. Annotation of reads is performed with respect to the mouse and rat transcriptomes, as well as annotated *de novo* contigs of CHO.

Contig annotation and selection pipeline. While knowledge-based contigs are by definition already assigned to their respective mouse transcripts, we used BLAST (27) with parameters optimized for more dissimilar sequence searches (word size of 16, a match score of 2, a mismatch score of −3, as well as gap open and gap extend penalties of −5 and −2, respectively, as suggested by the NCBI) to identify similar Ensembl mouse transcripts for CHO *de novo* contigs that are longer than 50 bp (all other contigs are ignored).

The hits returned by BLAST were filtered for matches with significant E-values of smaller than $10E^{-7}$ and hits where BLAST high scoring segment pairs (HSPs) cover at least 60% of a contig. This criterion led in most cases to a single mouse gene, which was assigned to the CHO contig. In the case of more than one mouse sequence matching the contig with the specified criteria, we selected the best transcript with respect to contig coverage and sequence identity. Unspecific contigs, i.e. those matching more than five transcripts with a similar quality, were filtered out.

Contigs which could not be assigned to any mouse transcript at all may represent misguided assemblies, novel transcripts, splice variants or non-conserved regions (e.g. UTRs) of known transcripts. They were not used for gene expression profiling.

Final CHO assembly. Finally, all contigs assigned to a gene in any of the three assemblies, *de novo* and knowledge-based, were combined and filtered for redundant information by detecting overlaps between the contigs. Overlapping sequences were merged, and singleton contigs with no overlap with others were also retained in the final set of contigs for a gene.

Read mapping pipeline

Reads were mapped to three different sequence sets (mouse, rat and CHO contigs) in parallel. Then we combined the information of the mapped reads in the second step. As discussed above, we used the mouse genome as a reference dataset. Therefore, reads mapping to mouse transcripts or CHO transcriptome contigs were summarized for the respective mouse genes. Reads mapping to rat transcripts were projected to mouse genes via the mapping of orthologous genes between mouse and rat provided by Ensembl. Reads mapping to multiple genes, either within a single reference dataset or between different references were handled as suggested by Mortazavi *et al.* (5). They were assigned to the respective genes proportional to the expression level of the respective genes as measured by unique reads.

Expression profiling and biological interpretation of the results

For the analysis of differential expression of genes between butyrate-treated cells and control cells using NGS data, reads obtained for each gene in a lane were counted and RPKM values were computed as proposed by Mortazavi *et al.* (5). Fold changes were computed as the ratio of the RPKM values obtained for a gene. The significance of differential gene expression was computed using the SAGEBetaBin method (28), and only genes with an absolute fold change >1.4 and a SAGEBetaBin significance score <0.01 were used for further analyses steps.

For Affymetrix arrays, which contain more than 23 000 probes detecting 10 425 Ensembl mouse genes, CEL files were processed using R and Bioconductor (29). Normalization was done using the MAS5 normalization (which normalizes all chips to the same mean value), and fold changes were computed as the ratio of the normalized and $\log_2$-transformed signal intensities. The LIMMA package (30) was used to compute adjusted *P*-values for all comparisons. In analogy to NGS data, all probes assigned to known Ensembl genes with an absolute fold change >1.4 and a *P*-value <0.01 were used in further analyses steps.

Results were interpreted in the context of biological processes and functions, as well as networks and pathways through the use of Ingenuity Pathways Analysis (Ingenuity Systems Inc., Redwood City, CA, USA). For the analysis of gene set enrichment in functional categories Fisher's exact test was used.

RESULTS

Throughout the analysis, we made extensive use of the mouse genome as a reference dataset for CHO sequences, i.e. to provide gene names for CHO genes and, more important, functional annotations of specific genes. This approach has also been proposed by Wlaschin *et al.* (31) and was used for the annotation of the existing CHO microarray. All CHO sequences will be analysed with respect to the corresponding mouse genes and are assigned to their specific mouse orthologs as described below. It has to been noted that the CHO EST sequences have not been used throughout this study as the data are not publicly available. But in principle, our pipeline also supports the use of such data in read mapping and analysis, if available. The analysis workflow is summarized in Figure 1 and a free Java implementation of the complete pipeline is available at http://unoseq.sourceforge.net.

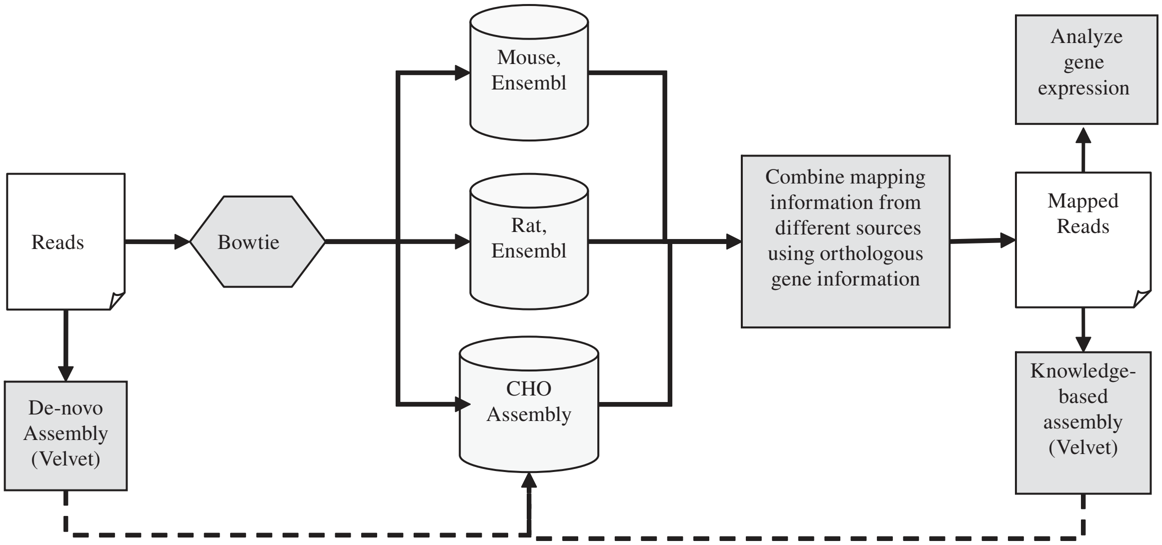
Reads
Bowtie
Mouse, Ensembl
Rat, Ensembl
CHO Assembly
Combine mapping information from different sources using orthologous gene information
Mapped Reads
Analyze gene expression
Knowledge-based assembly (Velvet)
De-novo Assembly (Velvet)

Figure 1. The bioinformatics workflow used for CHO data processing. In the first step, reads are assembled by two different strategies: (i) a *de novo* approach using Velvet starting with all reads and (ii) a knowledge-based approach using all reads mapping to a mouse gene (or the corresponding rat ortholog) again by applying Velvet. After computation and annotation of the final CHO transcriptome assembly, the reads that mapped against all three reference sequence datasets were used for gene expression profiling.

Table 1. Read mapping statistics for all 12 lanes sequenced on two flow cells

Lane	Reads	Mouse		Rat		CHO		All mapped
		All	Exclusive	All	Exclusive	All	Exclusive	
1	12201860	3569738	420713	3064690	235757	4168909	1597293	7369538
2	14627832	4390238	453707	3778402	281312	4996910	1840362	8866445
3	14382040	4306491	437651	3695549	272113	4936324	1814889	8692089
4	14853725	4375171	462080	3794065	285486	5128705	1930695	9004221
5	12824961	3462861	391391	2998615	233884	4117863	1578307	7363436
6	12622323	3473606	394837	2960651	242255	4226336	1735168	7140443
7	14063619	3497681	436575	3010972	267506	4369923	1856925	7508741
8	11349612	2919955	346243	2495134	216963	3685415	1553036	6163541
9	15799183	3965147	511692	3413354	298204	4984857	2147939	8666756
10	14629652	3411264	478117	2938477	265613	4315480	1891912	7564842
11	13611066	3743496	438070	3229333	279112	4571083	1903044	7801213
12	12166856	3061743	388744	2636835	236914	3813744	1627339	6557042

Lanes 1–5 are contributed by Flow cell 1, lanes 6–12 by Flow cell 2. Column 2 displays the total number of reads passing the quality filtering. The 'All' column shows the total number of reads mapping to a reference sequence dataset and the 'Exclusive' column shows the number of reads mapping only in the respective sequence set and not in the two others. Finally, the 'All Mapped' column shows the number of reads that can be recovered by the proposed mapping strategy.

Sequencing results

In total, we sequenced 12 lanes of CHO mRNA originating from three different cell cultures (control without butyrate, 0.5 mM butyrate as well as 1.0 mM butyrate) and three time points (Days 0, 6 and 8). In the control cell line, we sequenced two lanes for Day 0, three lanes for Day 6 and one lane for Day 8. From the 0.5 mM butyrate culture, one lane from each of the three time points were sequenced, while for 1.0 mM butyrate-treated culture two lanes for Day 6 and one lane for Day 8 were analysed. Each sample was sequenced on a single lane of a flow cell. Sequencing resulted in 11–15 mio. 36 bp reads per lane (Table 1) passing Illumina's quality filter with a total of 173 Mio. reads and 5.9 GB of CHO transcript sequence data. The results from Day 6, 18 h after butyrate addition, are the most interesting from a biopharmaceutical point of view, as the cellular effects leading to a higher cell-specific productivity of butyrate-treated cultures are likely to be detectable while cell viability is not yet significantly reduced by butyrate treatment. We, therefore, sequenced the control samples and 1.0 mM butyrate samples of Day 6 in technical replicates of three and two, respectively, in order to allow for a reliable estimation of the variance in read counts introduced by technical effects. Sequencing of different time points was performed to gather transcript data also on genes which are expressed only at specific days of the cultivation process. Reads were then used for transcriptome assembly and gene expression analysis. The bioinformatics workflow is summarized in Figure 1, and the results of the single steps of the workflow (assembly and mapping) will be described in the following. Short read data has been deposited in NCBI's Short Read Archive under project id SRA010967.

CHO transcriptome assembly: extending the CHO sequence space

Reads across the entire length of the transcripts present in the samples were obtained. In transcripts which are covered multiple times by read data, reads are likely to overlap and, therefore, can be assembled into longer contigs using short read assembly algorithms.

Table 2. Statistics on the contigs obtained by the two *de novo* assemblies

Description	Flow cell 1	Flow cell 2
Number of contigs (contig length >50 nt)	98 660	124 339
Average contig length	230	227
No significant BLAST hits in mouse	45 857	69 072
No significant BLAST hits in mouse (contig length >100 nt)	22 104	34 110
Number of significant BLAST hits in mouse (E-value $<10E^{-7}$)	52 803	55 267
Matching unique mouse genes	48 265	50 416
Distinct Ensembl genes mapped by contigs	9363	9762

Contigs were searched against the mouse Ensembl transcriptome with BLAST (as described in the 'Materials and Methods' section) and a strict E-value threshold of smaller than $10E^{-7}$ was used.

We assembled the reads obtained from the 12 lanes using two different assembly strategies as described in the 'Materials and Methods' section. This resulted in two *de novo* assemblies (one for each flow cell) and one knowledge-based assembly. The final CHO transcriptome assembly was then computed by merging all contigs assigned to a gene and resolving overlaps between the contigs.

The results of the two *de novo* assemblies are summarized in Table 2. We used BLAST (27) to assign contigs to the Ensembl mouse transcriptome. Two criteria were used to filter matches. We applied a strict E-value threshold of $10E^{-7}$ and demanded for at least 60% of the contig sequence to overlap with the mouse transcript. By this means, 53% (52 803) of the *de novo* contigs in Flow cell 1 and 44% (55 267) of the contigs in Flow cell 2 could be assigned to mouse orthologs. Out of those, >90% could be assigned to an unique gene in mouse and were, therefore, retained in the dataset. About 50% of the

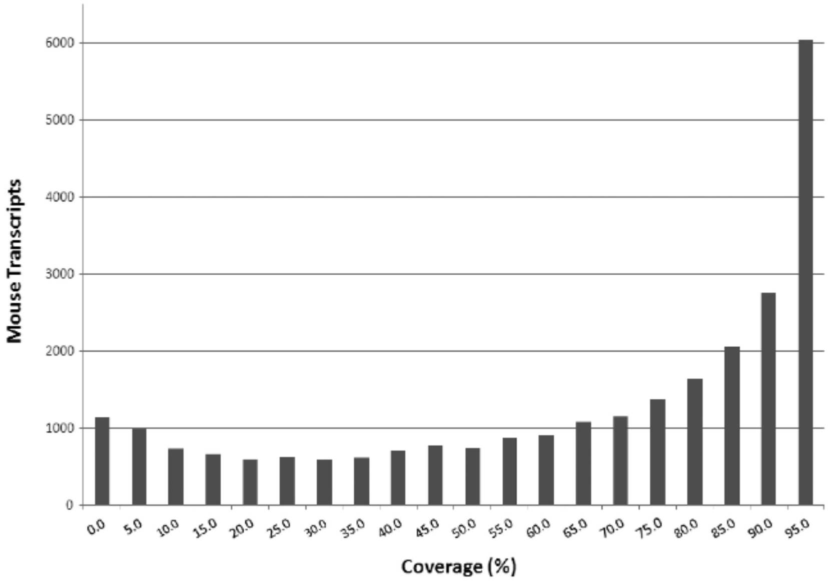

Figure 2. Coverage of mouse Ensembl transcripts by CHO contig sequences. More than 6000 transcripts can be expected to be nearly complete with a coverage of >95% with respect to mouse. The average mouse transcript coverage is 66.9%.

contigs that did not have any significant hit in the mouse transcriptome were smaller than 100 nt. Extending the reference sequence dataset to the complete RefSeq database allowed the assignment of 7954 additional contigs (14% of the contigs with no hits in the mouse Ensembl transcriptome) to unique RefSeq sequences which suffice the criteria defined above. These contigs were assigned to sequences from rat (2812 contigs), mouse RefSeq only sequences (1345 contigs), *Schistosoma mansoni* (930 contigs), human (919 contigs), *Macaca mulatta* (465 contigs), chimpanzee (419 contigs) and some other organisms.

In contrast to the *de novo* assembly strategy, for the knowledge-based assembly all reads mapping to a specific Ensembl mouse gene (or its corresponding rat ortholog) in any of the 12 lanes were collected and then assembled. This resulted in 93 016 contigs with an average length of 272 nt. As expected, these contigs were longer on average than contigs obtained from the *de novo* assembly; they represented sequences for 13 013 different mouse Ensembl genes.

Our final assembly was computed as a combination of *de novo* and knowledge-based assemblies of the CHO transcriptome and consisted of 92 272 contigs. These were assigned to 13 375 mouse Ensembl genes. The average length of the contigs could be increased to 352 bp by combining overlapping contigs. Almost 8000 contigs were >1000 nt in length with the largest ones having a length of >12 000 nt. They represent mRNAs of the *Protocadherin Fat 1* gene and the Serine/threonine-protein kinase *SMG1*.

Contigs were then aligned to the Ensembl mouse transcriptome using standard sequence alignment in order to estimate the completeness of the CHO contigs with respect to mouse transcripts. As shown in Figure 2, >6000 reference transcripts are almost completely covered by CHO sequence (coverage >95%), and therefore are likely to be also nearly complete for CHO. The average transcript coverage is 66.9%.

While the CHO Affymetrix microarray measures expression levels for 10 425 genes, at least 13 375 genes are detectable by NGS as they lead to assembled contigs. Additionally, lower abundance genes with orthologs in mouse and rat can be detected by reads mapping directly to mouse or rat transcripts. A comparison of the genes present on the chip and the genes present in the CHO assembly (Figure 3) shows that 8404 genes are detectable on both platforms, while 4971 genes are obviously expressed in the CHO cell line being analysed, but escaping detection on the chip.

By using this thorough pre-processing and assembly strategy for the read data, we could generate a significant amount of sequence information for CHO without any prior information on the CHO transcriptome. Additionally, as highly expressed genes lead to many reads which in turn can likely be assembled to contigs, we are able to profile exactly those genes that are really present in a specific cell line or under a specific treatment.

As shown above, longer contig sequences can be reliably assigned to orthologous genes in mouse (or other organisms) using BLAST, even in cases where less conserved parts including mutations, insertions and deletions exist. Since they represent the true CHO sequence of a transcript (including many of the CHO-specific insertions, deletions and mutations), reads originating from CHO are likely to fit better to CHO contigs than to transcriptomes of related organisms like mouse and rat. This is especially important, as short read mapping algorithms allow only for a limited number of mutations (to guarantee specificity of the mapping) and in most cases require non-gapped matches of the reads to the reference sequence. Reads originating from regions with a larger variability in CHO compared with mouse and rat can, therefore, only be detected knowing the true CHO sequence. In the following

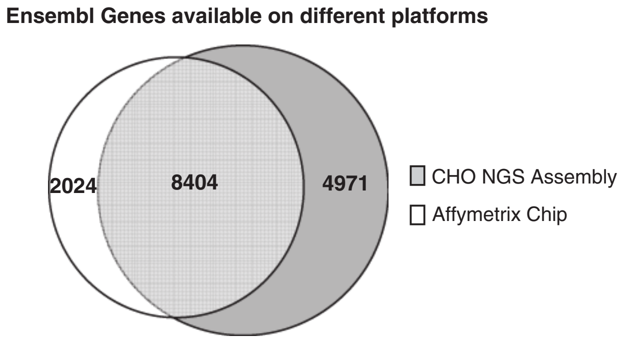

Figure 3. Overlap of Ensembl genes detectable on the custom Affymetrix CHO microarray and detectable using the CHO assembly. Please note that not all genes detectable by NGS during expression profiling are contained in the assembly, e.g. because they are expressed at too low levels. In those cases orthologous to mouse or rat can be detected by the mapping strategy and allow for profiling the expression of the corresponding CHO genes.

section, the CHO assembly proves to be very useful and allows the recovery of many reads that do not map to the transcriptomes of related organisms.

Making the most out of read data: read mapping pipeline

All reads have been mapped to three different reference datasets, namely mouse and rat (those mappings have already been used to compute the knowledge-based assembly) and to the final CHO transcriptome assembly in order to recover as many reads as possible and identify their genomic origin. We noted that adding the human transcriptome as a fourth dataset did not improve the mapping statistics, and therefore has not been used for further steps (data not shown). Based on the read mapping against the CHO transcriptome assembly, we estimated the sequencing error rate to be ~0.8% per base indicating a very high sequence quality of the short reads used in our experiment. More than 90% of the read map either perfectly to the reference sequence set or have at most one mismatch. For more details see Supplementary Table S1.

About 60% of all reads obtained in a lane could be assigned to at least one sequence in one of the reference datasets and were recovered for gene expression profiling (Table 1). The majority of mapped reads (~70%) map to more than one reference sequence dataset at the same time. In more than 90% of those cases, a single mouse gene (or in the case of a mapping to the rat transcriptome, the orthologous rat gene) was identified showing that the mapping of reads across the different species is highly consistent.

Finally, the statistics showed that mapping reads to a single reference sequence dataset is less powerful than the combination of all three datasets. This proposed mapping strategy can greatly help to recover the origin of as many reads as possible. As indicated, 30% of reads map exclusively to one reference dataset. The CHO assembly dataset contributes the majority of those reads, ~70% of the exclusive reads are mapped to CHO sequences showing the importance of performing a thorough pre-processing of the data for read mapping. About 55% of all mapped reads and 35% of all reads in each lane would have been recovered if only the assembled contigs were used. Since those contigs also contain information on reference transcripts in terms of the knowledge-based assembly, this number represents an upper limit of the reads which would be recovered without any information on genomes from related organisms.

Overall, the identity of a significant amount of reads could be determined. Those were used subsequently to perform a reliable, in-depth expression profiling of CHO cells undergoing sodium butyrate treatment.

Effects of butyrate treatment on cultivated CHO cells

In order to gain a deeper understanding of the effects of butyrate treatment in CHO cells, we analysed differentially expressed genes detected by mRNA sequencing in a controlled CHO cultivation process.

The effects of butyrate on the CHO cell culture are summarized in Figure 4. Prior to butyrate addition at cultivation Day 5.25, the viable cell number, the cell viability and the recombinant product concentrations were comparable across the three cultures, a fact which can be attributed to the controlled process conditions. Addition of butyrate to the CHO fed batch process resulted in a butyrate concentration-dependent decrease of the viable cell number from Day 6 onwards, a decline of cell viability was apparent from Day 7. Whereas the 1.0 mM butyrate treatment led to a decrease of the final product concentration at Day 10 compared with the control culture (grown under the same conditions but without butyrate addition on Day 5.25); a butyrate concentration of 0.5 mM in the cultivation medium slightly increased the final product concentration by ~5%. Since the viable cell number decreased at the same time, this translates into an increase of the cell-specific productivity of about 14% at Day 10 (compared with control).

In addition to NGS, expression profiling was also carried out using the custom Affymetrix CHO array in order to compare these two platforms. Ingenuity Pathway Analysis was used to analyse differentially expressed genes and to enrich affected processes with additional information from the literature.

As discussed above, we focused on the comparison between the control group (no butyrate treatment) and the cells treated with butyrate at 1.0 mM at Day 6 of the experiment (18 h after butyrate addition). Technical replicates (three lanes for the control group and two lanes for butyrate-treated cells) sequenced for those two samples were used as a reliable estimation of the differences between the cell lines due to random sampling effects and were included in the computation of the statistical significance of differences detected between the two samples.

In this comparison, 1785 genes were identified as being deregulated by an absolute fold change of at least 1.4 (and a significance value <0.01) in the NGS data. One thousand and thirty-seven genes were up-regulated in the control group while 748 genes were down-regulated under butyrate treatment. In comparison, using Affymetrix microarrays 595 genes were identified as being

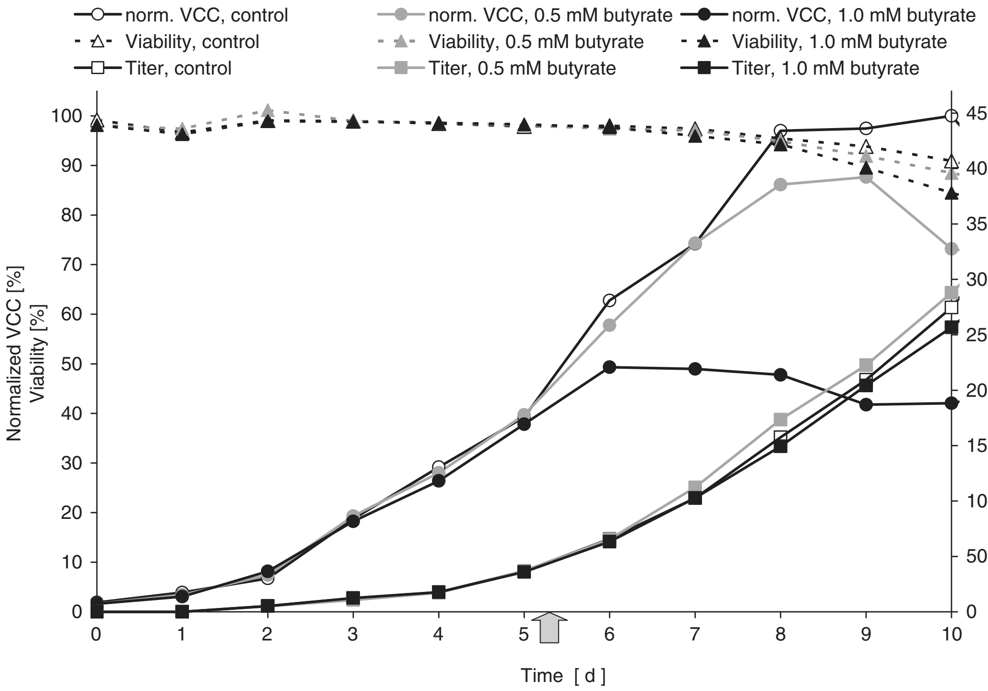

Figure 4. Cultivation data of recombinant CHO fed batch process. Normalized viable cell concentration (VCC; normalization to maximal VCC of control process without butyrate addition), cell viability and recombinant IgG product concentration are shown. Butyrate was added to the culture at $d = 5.25$ as indicated by the arrow. Sampling times for RNA extraction were $d = 0$, 4, 6 and 8.

differentially regulated (415 up and 180 down). The overlap between genes identified by NGS and by the CHO chip was large with 57% for the down-regulated genes (103 genes) and 55% of the up-regulated genes (231 genes). A significant amount of 752 deregulated genes could only be detected by NGS as they are not measured by a probe on the Affymetrix chip. Similar overlaps between genes detected by NGS and Affymetrix have also been described by Tang *et al.* (32).

The identified genes were involved in several biological functions, and many functional categories were significantly affected according to an over-representation analysis using Fisher's exact test. Processes identified by NGS along with the Affymetrix array included the regulation of the cell cycle, cellular growth and proliferation, DNA replication, recombination and repair, cellular assembly and organization, cellular movement, several metabolic pathways (like purine and pyrimidine metabolism) as well as other processes. Overall, those results were consistent with other published data on genes and functional categories regulated under butyrate treatment (20–22,33).

As shown in Figure 5a, NGS detected more significantly deregulated genes in those cellular processes known to be influenced by sodium butyrate than the Affymetrix chip. For example, with NGS we identified 197 genes involved in the regulation of the cell cycle in contrast to 115 genes detected on the chip. The majority of those additionally identified genes (53 in the specific example and 65% on average across all categories) originated from genes, which are not represented by a probe on the Affymetrix chip. For genes which were detected as significantly deregulated on both platforms, the correlation of their fold changes was high (Pearson correlation coefficient: 0.87, see also Supplementary Figure S1) indicating a good consistency of the expression measurements between NGS and Affymetrix chips which has also been described by Marioni *et al.* (34). As shown in Supplementary Figures S2 and S3 genes, which are measured as significantly deregulated in a platform usually showed the same direction of regulation (though not necessarily statistically significant) in the other platform. This indicates that many of the additional genes only detected by NGS as being significant, although they are measured on the chip, are due to the stringent threshold criteria applied to filter for significantly deregulated genes. Some of them may also have been identified because of the higher accuracy of NGS in measuring absolute expression levels as described by Fu *et al.* (35) and Marioni *et al.* (34).

Overall, while both technologies identify the same cellular processes as being regulated, NGS can provide a more detailed picture mainly due to the fact that the technology allows for an unbiased look into the transcriptome which is not restricted to those genes spotted on the Affymetrix chip.

In the following, we will analyse selected pathways and associated genes identified from the NGS data. A list with log ratios and *P*-values for all genes can be found in the Supplementary Data.

Regulation of cell-cycle genes. As shown in Figure 5a, 197 genes associated with the cell cycle (*P*-value: up to $6.8E^{-16}$) were differentially regulated between the control group and butyrate-treated cells. Below, we will describe two examples that—like many of theses genes—are involved in checkpoint control pathways (Figure 5b and c).

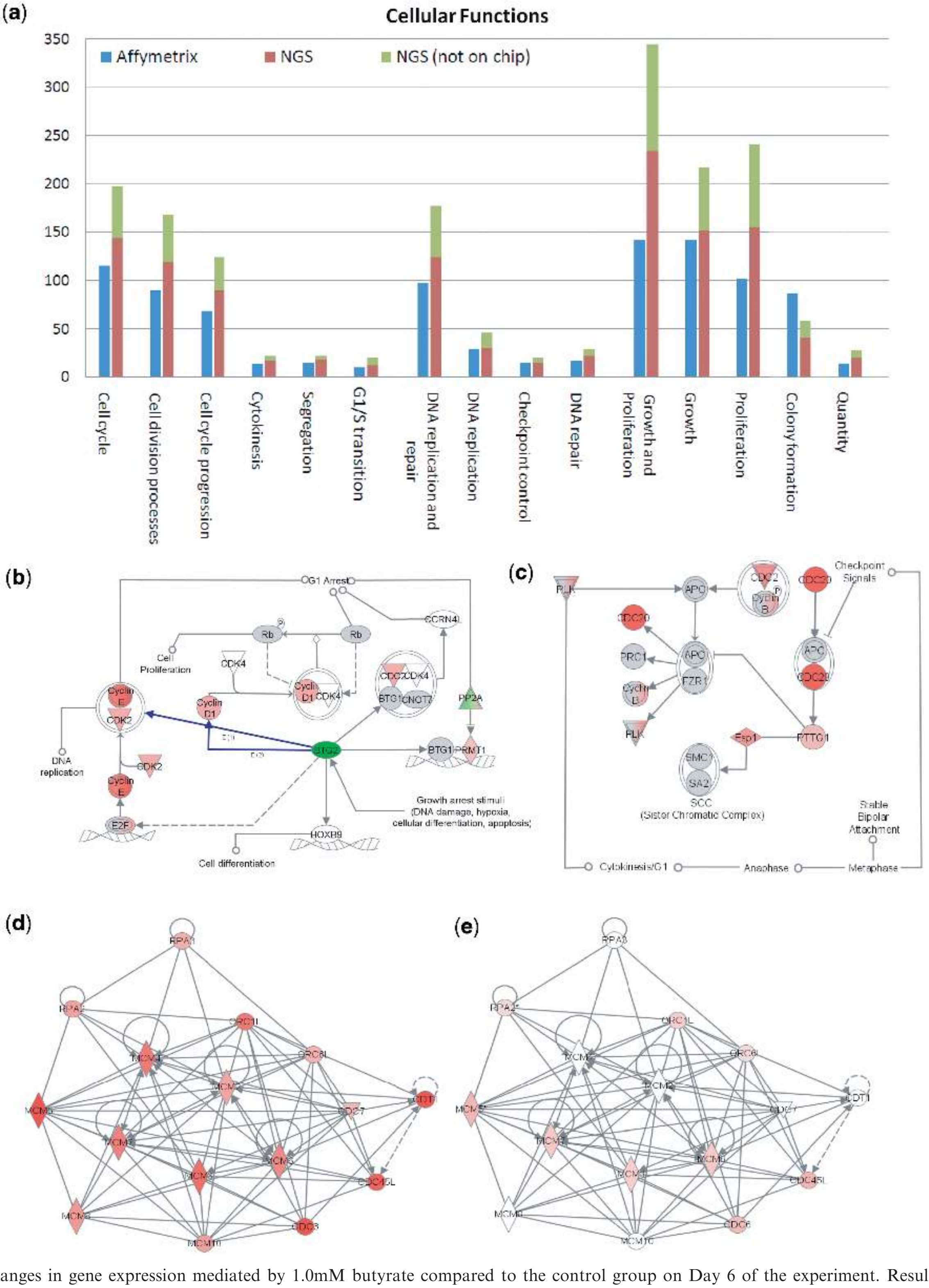

Figure 5. Changes in gene expression mediated by 1.0mM butyrate compared to the control group on Day 6 of the experiment. Results from gene expression analysis of cells treated with 1.0mM butyrate compared to the control on Day 6 of the experiment (18 h after butyrate addition). In the networks shown, genes marked in red are down-regulated in butyrate-treated cells whereas genes shown in green are up-regulated. Pathways and networks are taken from Ingenuity Pathway Analysis. (**a**) Top three functional categories and number of regulated genes detected by NGS and by Affymetrix, indicating that NGS provides a more detailed look into transcriptional changes mediated by butyrate. (**b**) and (**c**) Regulation of cell cycle genes through butyrate. (**b**) Up-regulation of BTG2 is known to lead to a down-regulation of the Cyclin E/CDK2 complex as well as Cyclin D1 (blue arrows). (**c**) Pathway of genes involved in mitotic check point control. (**d**) and (**e**) Example for the high consistency of differential gene expression detected by CHO NGS shown by the down-regulation of MCM genes at Day 6 in cells treated with 1.0mM butyrate. The MCM complex and associated genes play a role in the initiation and the elongation phases of eukaryotic DNA replication, specifically the formation and the elongation of the replication fork. (**d**) shows down-regulated genes as detected by the NGS approach while part (**e**) shows deregulated genes detected by the Affymetrix CHO chip. Several genes (MCM2, MCM4, MCM8 and CDT1) are not measured on the chip but fit well into the complete down-regulation of the complex.

In the cell cycle pathway shown in Figure 5b, *BTG2* was found to be up-regulated in butyrate-treated cells and may play an important role in butyrate-mediated cell cycle arrest. *BTG2* is activated by a number of different stimuli, including DNA damage or cellular stress and has been described to lead to a down-regulation of the *Cyclin E/CDK2* complex as well as *Cyclin D1* (36) that are essential in the progression of the cell cycle from G1 to the S phase, which is in agreement with our results. Among the proteins belonging to the cyclin family, which are responsible to regulate cyclin-dependent kinases, many members are differentially regulated after butyrate addition compared with the control. Examples for down-regulated cyclins are *CCNA2*, which promotes G1/S or G2/M transition through activation of *CDC2* or *CDK2*, *cyclin B1* and *B2*, *cyclin D1*, as well as *cyclins E1*, *E2* and *F*. *Cyclin G2,* which is known to increase cell cycle arrest is up-regulated as are *cyclin H* and *I*. *CDK2* and *CDK6* are also down-regulated under butyrate treatment, as well as many members of the CDC group of proteins like *CDC2*, which is part of the M phase promoting factor (MPF) complex, and others.

The almost 4-fold up-regulation of *TP53INP1* (tumor protein p53 inducible nuclear protein 1) and *SMAD3* are additional indications of the G1 phase arrest of butyrate-treated cells. *TP53INP1*, which has been shown to lead to cell cycle arrest in G1 (37), is functionally associated with *p73* to regulate cell cycle progression and apoptosis, independently from *TRP53* (which is down-regulated under butyrate treatment). *SMAD3* is the key mediator of TGF-β signaling and has also been shown to increase cell cycle arrest (38).

In addition to the cell cycle control genes, 22 genes (*P*-value: $2.2E^{-8}$) were involved in cytokinesis. Twenty were down-regulated in butyrate-treated cells including the concerted down-regulation of *Aurora kinases A* and *B* (which associate with microtubules during chromosome movement and segregation), as well as several members of the KIF family (kinesin superfamily of microtubule-associated motors) proteins which facilitate the movement of chromosomes during cell division.

Regulation of DNA replication, recombination and repair genes. Many genes regulated by butyrate were involved in DNA replication, recombination and repair. DNA replication was significantly affected (*P*-value: $6.9E^{-11}$) and contained 46 genes, 45 of which were down-regulated. The most prominent complex in this network, which is shown in Figure 5d and e, corresponded to a highly connected component containing the mini-chromosome maintenance complex (MCM) and associated genes like *CDC6* and *CDC7*, members of the origin recognition complex (*ORC1L* and *ORC6L*), *CDT1* and *RPA2* and *RPA3*. In comparison, the CHO chip identified merely nine out of the 15 genes in this complex as being down-regulated, but could not detect the regulation of the six other genes as they are not represented by a probe on the chip. The only gene which was up-regulated in this network was *RAD17*. *RAD17* is known to be involved in DNA damage recognition and, after being phosphorylated by ATR, it can induce the arrest of cells in the G2 phase.

Furthermore, 22 genes involved in the segregation of sister chromatids and chromosomes (*P*-value: $5.5E^{-8}$), 11 genes involved in the orientation and alignment of chromosomes (*P*-value: $2.5E^{-7}$), as well as 14 genes involved in the formation of the mitotic spindle (*P*-value: $6.8E^{-6}$) like *NDE1*, which is important for microtubule organization, were down-regulated under butyrate treatment.

All the above-mentioned genes are essential components of the eukaryotic DNA replication apparatus, and their concerted down-regulation after butyrate addition indicates that DNA replication prior to mitosis as well as important steps during mitosis are inhibited by butyrate.

Changes in the extracellular space. Several other differentially expressed genes affect the composition of the extracellular matrix or represent growth factors and chemokines. Several members of the matrix-metalloprotease family which are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction and tissue remodeling were found to be down-regulated in the treated samples (*MMP3*, *MMP9*, *MMP14*, *MMP19* and *MMP23B*). In correspondence with this finding were changes in the expression of collagens, where *Col1A1* was found to be >5-fold up-regulated in contrast to *Col4A5*, *Col5A2*, *Col6A1* and *Col8A1*, which were all down-regulated.

Besides changes in the extracellular matrix, we found several growth factors like CTGF and NGF to be up- or in the case of *FGF13* to be down-regulated.

TGF-β signaling pathway and other regulated processes. We also observed a down-regulation of the cytokine TGF-β and a reduction of its signalling by the down-regulation of well-characterized downstream targets like *SEPRINE-1*, *TGIF*, *JUN* and *JUNB* after butyrate treatment. Furthermore, TGF-β may be involved in remodeling of the extracellular space (as discussed above) by the regulation of metalloproteases and collagens, which is in line with the above-mentioned findings for butyrate-treated CHO cells, and therefore strongly indicate a role of butyrate treatment onto the TGF-β signaling pathway.

In addition to the processes and pathways discussed above, several metabolic pathways were influenced by butyrate. These included genes involved in pyrimidine and purine, as well as lipid metabolism. Again we found that NGS data can identify many more genes associated with those processes than the chip platform as shown in the supplementary Figure S4.

DISCUSSION

In recent years, expression profiling has mainly focussed on organisms with well-characterized genomes for which well-established chip platforms were available. However, from expression profiling of organisms, where only fragmented or no information on the corresponding genome or transcriptome is known (hence no chip platforms are currently available) and that play an important role with respect to biotechnological and pharmaceutical applications, relevant new insights can be expected.

In this study, we applied NGS to the problem of expression profiling without a reference genome using the example of CHO cells undergoing butyrate treatment. CHO cells lines are highly relevant for the production of biopharmaceuticals such as therapeutic antibodies.

We demonstrated that, by using a thorough preprocessing of the read data combined with an advanced mapping strategy, expression profiling in CHO cells is possible using NGS at a yet unknown resolution. By applying two different assembly strategies, i.e. *de novo* assembly of the reads without prior knowledge and a knowledge-based strategy, which utilizes reference sequences from mouse and rat, we could generate a significant amount of sequence information on the Chinese hamster transcriptome. Our assembly strategy could contribute partial transcript sequences for >13 000 CHO genes. More than 6000 of those transcript sequences are likely to be complete as they cover >95% of the orthologous mouse transcript mRNAs. On average, reference transcripts in mouse are covered by 66.9% with CHO sequences indicating that for expressed genes in the samples a large transcript coverage and a huge amount of sequence information could be generated. This sequence information has the potential to improve expression profiling for the CHO model in the future. Moreover, this information comes at no additional time and cost as expression profiling and sequencing of the CHO transcriptome are performed in a single experiment, a clear advantage over the expensive generation of EST libraries.

While making extensive use of the mouse genome to annotate CHO genes via BLAST, we noted that some caution is required with respect to the assignment of CHO gene functions. Due to variable degrees of sequence homology, not all of the transcripts profiled may indeed perform comparable functions in both organisms. Hence, our approach of defining gene identity will certainly not replace a thorough annotation of a given transcriptome. It must be kept in mind, however, that this is a general problem of all (genome) sequencing and annotation projects and that data quantity and data quality across different genomes do not necessarily compare. Moreover, a growing conundrum is generated by the gap between data generation and annotation.

In order to recover the genomic origin of as many mRNA reads as possible, we applied a mapping strategy which makes use of closely related species (mouse and rat), as well as annotated contigs from the CHO transcriptome assembly. In total, about 60% of the reads sequenced could be assigned to genes and were used for gene expression profiling. In comparison, when mapping e.g. mouse transcriptome reads on the relatively well-annotated mouse genome, one can usually map 80% of the reads to exons and UTRs of known transcripts. The other 20% are likely to correspond to yet unknown transcripts or splice variants, novel exons or UTRs and RNA genes. In our study, in Chinese hamster no complete genome is available, UTRs are likely to be less-conserved and CHO contains genes that are not present in mouse or rat. Therefore, 60% of the reads which can be mapped appears to be reasonable.

Nevertheless, these data also point to a larger number of yet unknown genes and expressed elements in CHO for which no close orthologs in mouse or rat exist.

As shown, 60% of the reads sequenced per sample are still enough to perform a thorough gene expression profiling of CHO, as demonstrated for the example of butyrate treatment. We could identify many of the key regulators involved in butyrate-mediated cell cycle arrest in the G1 or G2 phase of the cell cycle and a consistent down-regulation of many genes taking part in DNA replication prior to and during mitosis. Furthermore, several other processes in the butyrate-treated cells seem to be altered, starting from the composition of the extracellular matrix and the secretion of growth factors to a change in the level of proteins involved in ion transport, ubiquitin-mediated protein degradation and cytokinesis. In this analysis, NGS showed clear advantages over the existing Affymetrix chip platform for CHO. Throughout all affected pathways and biological functions, NGS identified more differentially expressed genes. The majority of those genes was contributed by 5000 additional genes which can be detected by NGS and which are not spotted on the Affymetrix chip. Additionally, NGS may also overcome problems in probe design and sensitivity in the absence of the complete genome. Our analysis presented a yet unseen data quality that sharpened our understanding of genes regulated by butyrate and uncovered a number of processes and metabolic pathways which could only be detected as significantly affected by NGS.

Taken together, our results show that detailed profiling of changes in gene expression is possible using NGS even if the genome sequence of that organism is unavailable. This requires a more advanced bioinformatics analysis pipeline compared to a standard expression analysis. Data processing and analysis included transcriptome assembly, contig annotation, as well as the combination of information on related organisms and assembled sequences during read mapping. We showed that, once such a pipeline has been established, NGS is a powerful new tool to step into the transcriptome of genomically 'unknown' organisms for which expression profiling is impossible or extremely time consuming and expensive.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Boehringer Ingelheim Pharma GmbH & Co KG.

Conflict of interest statement. None declared.

REFERENCES

1. Kahvejian,A., Quackenbush,J. and Thompson,J.F. (2008) What would you do if you could sequence everything? *Nat. Biotechnol.*, **26**, 1125–1133.
2. Metzker,M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
3. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
4. Hillier,L.W., Reinke,V., Green,P., Hirst,M., Marra,M.A. and Waterston,R.H. (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.*, **19**, 657–666.
5. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
6. 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.
7. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
8. McKernan,K.J., Peckham,H.E., Costa,G.L., McLaughlin,S.F., Fu,Y., Tsung,E.F., Clouser,C.R., Duncan,C., Ichikawa,J.K., Lee,C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.
9. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
10. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
11. Aggarwal,S. (2007) What's fueling the biotech engine? *Nat. Biotechnol.*, **25**, 1097–1104.
12. Birch,J.R. and Racher,A.J. (2006) Antibody production. *Adv. Drug Deliv. Rev.*, **58**, 671–685.
13. Seth,G., Hossler,P., Yee,J.C. and Hu,W.S. (2006) Engineering cells for cell culture bioprocessing—physiological fundamentals. *Adv. Biochem. Eng. Biotechnol.*, **101**, 119–164.
14. Schaub,J., Clemens,C., Schorn,P., Hildebrandt,T., Rust,W., Mennerich,D., Kaufmann,H. and Schulz,T.W. (2010) CHO gene expression profiling in biopharmaceutical process analysis and design. *Biotechnol. Bioeng.*, **105**, 431–438.
15. Wlaschin,K.F., Nissom,P.M., Gatti,M.L., Ong,P.F., Arleen,S., Tan,K.S., Rink,A., Cham,B., Wong,K., Yap,M. *et al.* (2005) EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol. Bioeng.*, **91**, 592–606.
16. Yee,J.C., Wlaschin,K.F., Chuah,S.H., Nissom,P.M. and Hu,W.S. (2008) Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. *Biotechnol. Bioeng.*, **101**, 1359–1365.
17. De Leon,G.M., Wlaschin,K.F., Nissom,P.M., Yap,M. and Hu,W.S. (2007) Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. *J. Biosci. Bioeng.*, **103**, 82–91.
18. Choi,Y.H. (2006) Apoptosis of U937 human leukemic cells by sodium butyrate is associated with inhibition of telomerase activity. *Int. J. Oncol.*, **29**, 1207–1213.
19. Lu,R., Wang,X., Sun,D.F., Tian,X.Q., Zhao,S.L., Chen,Y.X. and Fang,J.Y. (2008) Folic acid and sodium butyrate prevent tumorigenesis in a mouse model of colorectal cancer. *Epigenetics.*, **3**, 330–335.
20. Li,R.W. and Li,C. (2006) Butyrate induces profound changes in gene expression related to multiple signal pathways in bovine kidney epithelial cells. *BMC Genomics*, **7**, 234.
21. Yee,J.C., De Leon,G.M., Philp,R.J., Yap,M. and Hu,W.S. (2008) Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol. Bioeng.*, **99**, 1186–1204.
22. Hendrick,V., Winnepenninckx,P., Abdelkafi,C., Vandeputte,O., Cherlet,M., Marique,T., Renemann,G., Loa,A., Kretzmer,G. and Werenne,J. (2001) Increased productivity of recombinant tissular plasminogen activator (t-PA) by butyrate and shift of temperature: a cell cycle phases analysis. *Cytotechnology*, **36**, 71–83.
23. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
24. Birzele,F., Kuffner,R., Meier,F., Oefinger,F., Potthast,C. and Zimmer,R. (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, **36**, D63–D68.
25. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
26. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
27. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Vencio,R.Z., Brentani,H., Patrao,D.F. and Pereira,C.A. (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics.*, **5**, 119.
29. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
30. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
31. Wlaschin,K.F. and Hu,W.S. (2007) A scaffold for the Chinese hamster genome. *Biotechnol. Bioeng.*, **98**, 429–439.
32. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
33. Kantardjieff,A., Jacob,N.M., Yee,J.C., Epstein,E., Kok,Y.J., Philp,R., Betenbaugh,M. and Hu,W.S. (2010) Transcriptome and proteome analysis of chinese hamster ovary cells under low temperature and butyrate treatment. *J. Biotechnol.*, **145**, 143–159.
34. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
35. Fu,X., Fu,N., Guo,S., Yan,Z., Xu,Y., Hu,H., Menzel,C., Chen,W., Li,Y., Zeng,R. *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.
36. Tirone,F. (2001) The gene PC3(TIS21/BTG2), prototype member of the PC3/BTG/TOB family: regulator in control of cell growth, differentiation, and DNA repair? *J. Cell Physiol.*, **187**, 155–165.
37. Tomasini,R., Seux,M., Nowak,J., Bontemps,C., Carrier,A., Dagorn,J.C., Pebusque,M.J., Iovanna,J.L. and Dusetti,N.J. (2005) TP53INP1 is a novel p73 target gene that induces cell cycle arrest and cell death by modulating p73 transcriptional activity. *Oncogene*, **24**, 8093–8104.
38. Nicolas,F.J., Lehmann,K., Warne,P.H., Hill,C.S. and Downward,J. (2003) Epithelial to mesenchymal transition in Madin-Darby canine kidney cells is accompanied by down-regulation of Smad3 expression, leading to resistance to transforming growth factor-beta-induced growth arrest. *J. Biol. Chem.*, **278**, 3251–3256.