Sequencing on the SOLiD 5500xl System – in-depth characterization of the GC bias

Simone Roeh, Peter Weber, Monika Rex-Haffner, Jan M. Deussing, Elisabeth B. Binder, and Mira Jakovcevski

Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany; Department of Stress Neurobiology and Neurogenetics, Munich, Germany; Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA

ABSTRACT
Different types of sequencing biases have been described and subsequently improved for a variety of sequencing systems, mostly focusing on the widely used Illumina systems. Similar studies are missing for the SOLiD 5500xl system, a sequencer which produced many data sets available to researchers today. Describing and understanding the bias is important to accurately interpret and integrate these published data in various ongoing research projects. We report a particularly strong GC bias for this sequencing system when analyzing a defined gDNA mix of 5 microbes with a wide range of different GC contents (20–72%) when comparing to the expected distribution and Illumina MiSeq data from the same DNA pool. Since we observed this bias already under PCR-free conditions, changing the PCR conditions during library preparation – a common strategy to handle bias in the Illumina system - was not relevant. Source of the bias appeared to be an uneven heat distribution during the SOLiD emulsion PCR (ePCR) - for enrichment of libraries prior loading – since ePCR in either small pouches or in 96-well plates improved the GC bias.

Sequencing of chromatin immunoprecipitated DNA (ChIP-seq) is a common approach in epigenetics. ChIP-seq of the mixed source histone mark H3K9ac (acetyl Histone H3 lysine 9), typically found on promoter regions and on gene bodies, including CpG islands, performed on a SOLiD 5500xl machine, resulted in major loss of reads at GC rich loci (GC content ≥ 62%), not explained by low sequencing depth. This was improved with adaptations of the ePCR.

Introduction
Next Generation Sequencing (NGS) techniques have become increasingly popular methods, since they are powerful tools for genomics and epigenomics research. Several different NGS systems have emerged in parallel using various approaches of library generation and performing of the actual sequencing procedure. Besides the currently most popular Illumina sequencing technology, the SOLiD System (Applied Biosystems) has been among the most frequently used NGS platforms. Data generated by both approaches is a highly valuable source for meta-analyses of any type. However, the specific weaknesses and strengths of the sequencing technique used, have to be taken into account for meaningful interpretation of the data.

One of the common problems of NGS techniques is the under- or over-representation of GC or AT rich sequences. These biases are often generated during library preparation, mostly when libraries undergo an upscale polymerase chain reaction (PCR), while PCR-free libraries are believed to be sequenced with significantly less to almost no bias. However, often an upscale PCR cannot be avoided for many applications from single cell sequencing to ChIP-seq (sequencing of DNA obtained from chromatin immunoprecipitation), due to low amounts of starting material. Obviously, concerning each specific application, the weakness of a particular NGS technology might have important consequences on the quality of particular data sets. For example, among the most frequently used techniques in epigenetics is genome-wide ChIP-
Figure 1. Characterization and improvement strategy of the GC sequencing bias on SOLiD 5500xl machines. (A) Theoretical distribution of reads (increasing GC content from left to right, visualized in a light-gray to dark-gray color gradient) overlaid with the actual (experimental) number of reads in the entire microbial gDNA mixture in relation to %GC content per fragment under different ePCR conditions, i.e., E80 ePCR pouch (dark-blue line), E20 ePCR pouch (medium-blue line), 96 well plate (contents of an E20 ePCR pouch distributed into 96 well plates, light-blue line), in comparison to an Illumina library (orange line). (B) % of shift of genomic content for the different microbes, plotted in relation to ePCR conditions and compared with Illumina sequencing results, Plasmodium falciparum (20% GC, violet line), Pictrophilus torridus (36% GC, light-blue line), Escherichia coli (51% GC, cyan line), Pseudomonas putida (62% GC, light-green line) and Micrococcus luteus (70% GC, dark-green line). (C-F) GC sequencing bias for the individual microbes. Panel shows theoretical (black lines) and experimental (colored lines) read frequencies for each of the microbes under the different sequencing conditions. (From left to right) highest to lowest GC content: M. luteus (dark green), P. putida (light green), E. coli (cyan), P. torridus (light blue) and P. falciparum (violet). (From top to bottom) (C) SOLiD E80 ePCR pouch (blue), (D) SOLiD E20 ePCR pouch (cyan), (E) SOLiD 96 well plate (light blue) and (F) Illumina MiSeq (orange).
seq of transcription factors or histone marks. Histone modifications are present in a wide range of genomic regions, depending on the particular modification and on the investigated cell type.\textsuperscript{9,10} The range from point source, via broad source to mixed source distributions among histone marks is associated with a distribution over a vast range of genomic elements that can be potentially challenging for sequencing.\textsuperscript{9} Especially, for ChIP-seq data or DNA methylation studies, quality depends on the amount of bias introduced during library preparation or sequencing, since both types of marks have been reported to be associated with challenging GC rich CpG islands.\textsuperscript{11,12}

So far, GC and AT biases have been well addressed and successfully reduced for the Illumina technologies by adapting PCR protocols, including changes to denaturation times and ramp rates, and using optimized polymerases for library amplification.\textsuperscript{3,13} In contrast, comparable approaches are missing for the SOLiD system. Thus, we were interested to investigate how strong the sequencing bias in the SOLiD system might be and to which extend this would affect ChIP-seq data. With the idea to characterize and improve the potential bias for the SOLiD machine and make it more suitable for applications such as ChIP-seq, we initially tested PCR-free library preparations for potential GC and other biases on this system.

In the present study, we observed a very strong GC bias in SOLiD sequencing data that exceeded previous estimations of the tentative sequencing bias in this system. We identified the emulsion PCR (ePCR) of the SOLiD sequencing system as the major source of GC bias. In response to that, we show approaches to reduce this bias by changing the ePCR conditions accordingly. Likewise, ChIP-seq of acetylated Histone H3 lysine 9 (H3K9ac), a mixed source histone mark, on a SOLiD 5500xl machine resulted in data almost void of reads at GC rich loci (GC content higher than 62%) while these regions were well covered by an Illumina sequencing system. ChIP-seq data on H3K9ac on the SOLiD was improved by optimized ePCR conditions and sufficient sequencing depth.

**Results**

**Characterization of GC bias on the SOLiD 5500xl system**

First, we aimed to characterize and potentially improve the sequencing bias of the SOLiD 5500xl machine. While there are several previous studies on how to counteract sequencing biases for the Illumina system, i.e. mostly by adaptions to the library construction protocols, no such attempts have been reported for the SOLiD library preparation workflow. Typically, these studies utilize mixes of small microbial genomes (1.5 to 24 MB, as opposed to human 3.2 GB\textsuperscript{3,6,13}). Such microbial gDNA mixes contain microorganisms with different GC content covering the whole spectrum of high, low and medium genomic GC content, allowing the characterization of the full range of bias of a protocol and its potential improvement through applied modifications.

In this study, we mixed 5 microbes with genomic GC contents of 20%, 36%, 51%, 62% and 72% that would allow for a good characterization of the sequencing performance. We build a PCR-free library from this gDNA mixture using the standard protocol and chemistry. Against the general view that the sequencing bias in a library should be minimized in the absence of an upscale PCR, but in line with our hypothesis, a strong bias was already present in the PCR-free library. Both, regions with very high AT content and to an even stronger extent those with high GC content were underrepresented in the SOLiD sequencing data (Fig. 1A–C). Comparing our SOLiD data to the expected GC distribution (calculated based on genomic sequence of the pooled microbes), we observed a shift in distribution for all microorganisms – with the strongest shifts for organisms with high GC and high AT and moderate to negligible shifts in organisms with moderate GC/AT content (see Table 1). Subsequently, the exact same microbial gDNA mixture was used to build a PCR-free library for the Illumina sequencing system (i.e., MiSeq) using standard Illumina reagents. Indeed, this library had a reduced sequencing bias when compared with the expected GC distribution and to the data of the PCR-free library sequenced on the SOLiD machine.

**Table 1. Summary of shift for the different microbes and ePCR conditions and for Illumina bridge PCR.**

| Microbe                  | GC content | SOLiD E80\textsuperscript{*} | SOLiD E20\textsuperscript{*} | SOLiD 96\textsuperscript{*} | Illumina |
|--------------------------|------------|-------------------------------|-------------------------------|-------------------------------|----------|
| Plasmodium falciparum    | 20%        | 46%                           | 39%                           | 38%                           | 21%      |
| Picrophilus torridus     | 36%        | 2%                            | 1%                            | 1%                            | 15%      |
| Escherichia coli         | 51%        | 24%                           | 18%                           | 17%                           | 16%      |
| Pseudomonas putida       | 62%        | 66%                           | 34%                           | 25%                           | 22%      |
| Micrococcus luteus       | 72%        | 96%                           | 86%                           | 74%                           | 23%      |

\textsuperscript{*}ePCR performed in an E80 or E20 pouch or in 96 well plates.
5500xl system (Fig. 1A, B and F). The shift for the same gDNA mix, sequenced on the Illumina MiSeq, was comparable low to moderate for all microbe genomes (Table 1). Interestingly, for Picrophilus torridus (with a moderate to low GC content of 36%), sequencing bias was significantly lower in the SOLiD library than for the Illumina library (2% versus 15% shift, Fig. 1A–C and F).

Since the SOLiD PCR-free library already presented with a strong bias, attempts to improve it by changing the upscale PCR (PCR protocol or polymerase) did not seem useful. The data rather suggested that the sequencing bias is largely introduced during the SOLiD emulsion PCR (ePCR), the equivalent to the Illumina bridge PCR (needed for cluster generation). Thus, we aimed to address this hypothesis by changing the ePCR conditions.

**Improving the SOLiD GC bias with modifications to the emulsion PCR (ePCR)**

Typically, the ePCR is performed in pouches of different volumes (from 10 mL to 120 mL, depending on the amount of libraries to be sequenced) in a bead amplifier, before loading the libraries onto the sequencer. Since we initially used a larger pouch of 80 mL for cluster generation/amplification of the libraries from the microbial gDNA mix, we hypothesized that a potentially uneven heat distribution in the pouch may have contributed to the observed bias.

To test this, we performed the emulsion PCR of the exact same microbial gDNA mix library, used throughout this study, in a 20 mL pouch and in 96 well plates. Indeed, both approaches improved the sequencing quality (Fig. 1A, B and D, E). Sequencing data with prior ePCR in a 20 mL pouch showed a reduction of the shift for all genomes compared with the 80 mL pouch ePCR condition, with most pronounced improvements for the GC rich genomes. The 96 well plate condition improved the SOLiD sequencing data further. We observed a reduction of the shift for all genomes compared with the 80 mL and 20 mL (except for the well represented Picrophilus torridus genome) pouch ePCR conditions (Fig. 1A–E, Table 1).

**H3K9ac ChIP-seq on the SOLiD 5500xl sequencer is prone to GC bias**

To determine to what extent this bias may hamper typical applications such as histone ChIP-seq, we first sequenced H3K9ac ChIP-DNA under standard conditions (ePCR in a 80 mL pouch). We deliberately selected a mixed source histone mark to get a better idea on how the bias might affect different genomic elements.

Sequencing of H3K9ac ChIP-DNA (prepared from adult mouse hippocampus) on the SOLiD 5500xl sequencer showed a strong enrichment compared with input controls (Fig. 2A, blue vs. gray tracks), indicative of reliable performance of the ChIP procedure. In line with published data on H3K9ac ChIP-seq, in e.g. embryonic stem cell (ESC) nuclei,14 we find H3K9ac signals not only restricted to the transcription start site (TSS) of genes (Fig. 2B), which is typical for mixed source distribution marks.9 We observed H3K9ac occupancies mostly on intronic regions, followed by intergenic regions, exons and promoter regions. In addition, we found occupancies in 5′UTRs, CpG islands, 3′UTRs and down-stream of the TSS.

Yet, as expected, we observed the strong GC bias in H3K9ac ChIP-seq as well. Similar to the GC bias present in our data from the microbial mix gDNA, respective regions with high and low GC content were most strongly affected. We noticed a complete absence of coverage for some loci, such as CpG islands with >62% GC content. Specifically, we observed a good coverage and highly elevated ChIP signal for the shore regions of CpG islands and a sudden decrease or complete loss of coverage for the islands themselves which typically have a higher GC content than their shores. These loci missing in the ChIP-seq data had no coverage in the input samples, strongly indicative of a sequence and not histone-mark dependent loss of coverage (Fig. 2A and D).

The peaks with the loss of coverage on the SOLiD sequencer were much better preserved when sequenced with the Illumina system MiSeq (Fig. 2A and D, orange tracks). Overall, compared with the H3K9ac ChIP-seq data generated with the Illumina sequencing technology, the data from the SOLiD 5500xl showed a loss of reads at several genomic elements, including exons and 5′UTRs. Specifically, CpG islands had a much lower coverage which is explained by the fact that loci with high GC contents (above 62%) were extremely underrepresented when sequenced on a SOLiD 5500xl system. In contrast, promoters, introns, 3′UTRs, and the shores of CpG islands were better preserved on the SOLiD 5500xl system.
Figure 2. H3K9ac ChIP-seq on SOLiD 5500xl vs. Illumina MiSeq. (A) Representative UCSC genome browser screenshots from H3K9ac ChIP-DNA sequenced on a SOLiD 5500xl (blue tracks) and an Illumina MiSeq (orange tracks). Note enrichment of ChIP-DNA tracks over input controls (gray tracks). Light blue boxes indicate gaps in the SOLiD 5500xl sequencing tracks, as compared with the Illumina sequencing tracks, typically, over CpG islands (green bars). (B) (Left) graph shows genomic elements with occupancy with the H3K9ac mark, plotted with corresponding number of peaks. (Right) dotted grid displays distribution of peaks over these genomic elements, including the information if peaks cover multiple genomic elements. (C) (Left side of the panel) Cartoon of experimental design for H3K9ac ChIP-seq with ePCR in 80 mL and 20 mL pouches (SOLiD) and with bridge PCR (Illumina). (Right side of the panel) Box-plots show sequencing coverage for H3K9ac on the SOLiD 5500xl (20 mL pouch, cyan and 80 mL pouch, blue) and on an Illumina platform (MiSeq, orange) for promoters, exons, introns, 3'UTRs, and 5'UTRs. (D) CpG island plots showcase coverage for the island and its shores, continuously 2kb up and downstream from the CpG island. Coverage is expressed in reads per million (RPM).
islands were covered to comparable extent by both sequencing systems (Fig. 2C, D).

Next, we were interested if our observation was common and generalizable. Therefore, we re-analyzed publically available H3K9ac ChIP-seq data (i.e., on mouse ESCs) from both sequencers. As expected, the Illumina and SOLiD sequencing data differed in their coverage comparable to our results. Likewise, we observed a loss of reads for promotor, exons, 5’UTRs, and the most pronounced loss of coverage for GpG islands (Fig. S1), suggesting that this underrepresentation, likely due to the GC bias in the SOLiD 5500xl system, can be generalized for data sets generated with this method.

**Improving H3K9ac ChIP-seq data**

Given that our modifications to the SOLiD ePCR improved the GC sequencing bias for the microbial gDNA mixture, we tested whether this would sufficiently improve our data on H3K9ac ChIP-DNA. Indeed, compared with the library, processed for ePCR in a big E80 pouch (80 mL), using a smaller E20 pouch (20 mL) in this enrichment step improved the data significantly. We primarily gained reads for the previously severely underrepresented GpG islands (Fig. 2D).

**Consideration on sequencing depth for H3K9ac**

Finally, we wondered if the lack of coverage could be partially attributed to insufficient sequencing depth which is also relevant for the interpretation of other published H3K9ac-seq data sets generated on the SOLiD 5500xl. In particular, this question is of high interest when sequencing or analyzing a histone mark with mixed source distribution such as H3K9ac which requires a higher sequencing depth than point source marks such as H3K4me3 to reach sufficient coverage.9

In the present study we sequenced with a depth of 40 Million reads per H3K9ac sample which exceeds the recommendation by ENCODE and those commonly found in the literature.9,15 Our sequencing depth reached saturation, indicating that we were sequencing in a sufficient range (Fig. S2B).

To test to which extent lower read numbers would affect H3K9ac data we applied first an in silico random down-sampling approach. Read numbers of the H3K9ac sequencing track were randomly removed, down to 20 Million, 10 Million and 5 Million reads. As expected, with every down-sampling step we lost actual coverage, affecting even the highly covered promoter regions at a sequencing depth as low as 5 Million reads. Also, signals on the gene body were lost with 20 Million reads. To evaluate this finding experimentally, we specifically tested latter loci that lost their coverage from 20 Million reads on by qPCR for their relative level of occupancy. These regions were compared with loci on the same gene that had H3K9ac signals under all levels of coverage, for a positive control, and to loci on the gene body and intergenic loci (in proximity to this gene) which did not show occupancies with 40 Million reads, as putative ‘negative’ controls. As predicted, amplification for the target loci was comparable to the positive control (Hap1) or even exceeded the value of the positive control locus (Meis2), suggesting that a sequencing depth of below 40 Million reads/per sample for H3K9ac ChIP-seq will lead to significant loss of data. Furthermore, we detected moderate qPCR signals for loci on the gene body and to some extend in intergenic regions (e.g., for Hap1), clearly indicating that H3K9ac ChIP-seq can benefit from increasing the sequencing depth beyond 40 Million reads per library (Fig. S2C). These intergenic regions might be highly relevant for studies into genetic and epigenetic mechanisms of neuropsychiatric disease mechanisms, since e.g., relevant GWAS risk loci often map to intergenic regions.16

**Discussion**

This study provides an in-depth characterization of the sequencing bias of the SOLiD 5500xl sequencer that was previously not well described and delivers profound insights into addressing biases in this system. We report that data generated on the SOLiD 5500xl sequencer is strongly impacted by an extensive GC bias. In contrast to theoretical assumptions, this bias was present in a PCR-free protocol and thus cannot be improved by changing the library preparation conditions such as changing the PCR-polymerase or the ramp rate of the thermocycler during upscale PCR as previously applied with good success for Illumina libraries.3,13

The presence of the bias in the PCR-free protocol led us to identify the emulsion-PCR step (ePCR, the SOLiD equivalent to the Illumina bridge PCR, required for cluster generation) as the major source of GC bias. In fact, even heat distribution in the emulsion
reaction cocktail has been reported as essential for ePCR applications. Thus, we aimed to improve the GC under-representation by changing the ePCR conditions. And indeed, both ways to improve the uniformity of thermal conditions, i.e., using a smaller ePCR-pouch and distributing the contents of a small ePCR-pouch into 96 well plates for running the ePCR reduced the GC bias significantly. Yet, our attempts to improve the sequencing performance for GC rich loci did not reach the level of on an Illumina machine, particularly at very high GC contents. The difference between the Illumina library and the SOLiD library enriched in an E20-pouch and a 96 well plate, respectively, was 12% and 3% for the Pseudomonas putida (62% GC content) genome, but 63% and 51% for Micrococcus luteus (72% GC content). It is noteworthy though that the observed bias was minimal for the P. torridus microbe (GC content of 36%) under all sequencing conditions on the SOLiD platform and far better than the shift observed on the Illumina system, suggesting the SOLiD as the ideal platform of choice for researchers dealing with organisms in a similar range of GC content.

It is conceivable that other factors than the evenness of heat distribution may be additional sources of the observed GC bias in the ePCR, such as the polymerase used or the ePCR protocol itself. While we already reduced the number of cycles in the ePCR by one third and a further reduction may lead to an incomplete enrichment of beads, additional approaches might consist of changes to ramp rate, denaturation and annealing length and time and temperatures, and the polymerase used for ePCR.

In a next step we demonstrate to which extent actual common sequencing applications are affected by the SOLiD sequencing bias and if they can be rescued sufficiently. We chose histone ChIP-seq for the mixed source distribution mark H3K9ac (Fig. S2A), as an example of an application that requires good coverage of multiple genomic elements such as introns, exons, promoter regions and CpG islands. We observed that H3K9ac ChIP-seq on a SOLiD 5500xl machine led to an extensive loss of coverage in regions with high GC content (> 62%) in comparison with samples run on an Illumina sequencer. In line with our data on the microbial gDNA mix, this bias was again most pronounced in GC rich loci. The bias was not due to an overall bad sequencing quality of H3K9ac ChIP-seq, which actually exceeded ENCODE standards and recommendations by recent literature (of 20–40 Million reads for ChIP-seq) in sequencing depth. Notably, sequencing depth reached saturation (Fig. S2B), and thus our experiments were clearly performed with bona fide sequencing depth for a mixed source distribution histone mark such as H3K9ac. Moreover, the bias was not associated with bad quality of our ChIP-DNA, e.g., caused by failure of the ChIP procedure or over-shearing of DNA, since our tracks showed strong enrichment of histone acetylation signals, the GC bias occurred in the input samples as well (Fig. 2A), and the phenomenon was also present in published data sets of other groups (Fig. S1). Interestingly, the latter data sets presented with a significant, but less pronounced bias than our ChIP-seq data. This might be due to the wider distribution of H3K9ac in adult brain (our study), e.g., we observed broader and higher peaks than in embryonic stem cells. Notwithstanding, our bias may be exceedingly strong, because we sequenced sufficiently deep and thus more peaks (on the Illumina sequencer) or the absence of them (in the SOLiD system) were detected. This view is supported by our down-sampling data in conjunction with ChIP-qPCR, indicating that 5 million reads hardly suffice to detect even all major reads at the promoter regions and a sequencing depth of 20 million reads may still miss loci with H3K9ac signal on the gene body.

In conclusion, Illumina platforms seem to be better suited for applications that require sequencing of GC rich loci such as histone ChIP-seq and presumable DNA methylation studies with focus on CpG islands. Comparable sequencing coverage in the Illumina and SOLiD systems can be reached at balanced levels of GC content (e.g., E. coli, GC content 51%), and better coverage in the SOLiD sequencing libraries can be obtained for lower GC contents (e.g., P. torridus, GC content 36%), with and without improved ePCR.

Above all, it is essential to know the extent of the GC bias in the SOLiD 5500xl system for better interpretation of data previously generated on this machine. These data can still be valuable for diverse meta-analysis and may be improved by pooling samples during analysis to reach higher coverage and detect all potential peaks in ChIP samples. Findings from this study are also relevant for the interpretation of data, obtained with other sequencing systems such as Polony Sequencing, Roche 454 or Ion Torrent using ePCR for enrichment.
Material and methods

gDNA isolation and preparation (microbial gDNA)

All microbes were grown in appropriate media under specific conditions required for each of the organisms and thereafter pelleted and stored at $-20^\circ$C until use. gDNA was isolated using Tris-HCl lysis buffer (1% SDS, pH 8.0) and TE-buffered Roti-Phenol/Chloroform/Isoamyl alcohol (25:24:1, pH 7.5–8.0, Carl Roth, #A156.2). Samples were precipitated with ethanol and purified with QIAquick spin columns (Qiagen, #28104). Contaminating RNA (which could interfere with the quantification of gDNA) was digested with RNase A. Aliquots of each sample were run on an agarose gel to check size and purity of the respective gDNA. Concentration of samples was measured with the Qubit Fluorometer (Life Technologies, #Q33216). Thereafter, the 5 different microbes were pooled in the following proportion: Picrophilus torridus: 100%, Plasmodium falciparum: 140%, Pseudomonas putida: 60%, Escherichia coli: 80% and Micrococcus luteus: 100%. All microbes were grown in appropriate media under standard conditions for each of the organisms specified in the manufacturers manual. Thereafter samples were loaded onto the SOLiD 5500 xl machine (Applied Biosystems, # 4460730).

Library preparation and sequencing

1/ Library preparation for SOLiD sequencing

Libraries were built from ChIP DNA, input DNA or gDNA (microbial gDNA mixture) using the fragment library preparation kit for 5500 series SOLiD systems (Applied Biosystems by Life Technologies, #4464412) according to the manufacturer’s instructions. ChIP and input libraries were amplified on a Gene Amp PCR System 9700 (Applied Biosystems, #N8050200) with standard settings (ramp rate 5°C/s) and SOLiD standard reagents (AmpliTaq Gold polymerase, Applied Biosystems, #N8080241). The library from microbial gDNA was prepared amplification free. All libraries were quantified on a Qubit Fluorometer and checked for expected size on a Bioanalyzer (Agilent Technologies, #G2939AA).

For bead emulsion PCR (ePCR) the SOLiD EZ Bead E80 System (Applied Biosystems #4472999) was initially used. Alternatively, E20 System pouches (Applied Biosystems, #4453094) were used instead. Pre-set cycling conditions, used on the bead amplifier (ePCR, Applied Biosystems, #4448419) were as follows: 95°C for 350 sec, 60°C for 60 sec and 75°C for 75 sec, followed by 60 cycles of 96°C for 65 sec, 60°C for 60 sec and 75°C for 75 sec, and a final step of 75°C for 420 sec, 50°C for 120 sec and 30°C for 12 sec, each. In a third approach the content of a E20 bag was distributed to 96 well plates and run on a thermocycler (ABI 9700) using the following conditions: initial denaturation (95°C for 5 min), followed by 40 cycles of denaturation (93°C for 15 sec), annealing (62°C for 30 sec) and extension (72°C for 75 sec), followed by final heating (72°C for 7 min). ePCR was cleaned up and enriched to a concentration of approximately 1.5 million beads/μL, according to the color scale provided in the manufactures manual. Thereafter samples were loaded onto the SOLiD 5500 xl machine (Applied Biosystems, # 4460730).

2/ Library preparation for Illumina sequencing

Libraries were prepared using either the Illumina TruSeq ChIP Sample Preparation kit for ChIP libraries (Illumina, #IP-202–1012) or the TruSeq DNA PCR-free sample preparation kit (Illumina, #FC-121–3001) for library preparation from the microbial gDNA mixture. The ChIP DNA library was amplified by PCR (Biorad T100 Thermal Cycler, #1861096). The following PCR conditions were used: denaturation for 1 min at 95°C, 15 cycles of denaturation (at 95°C for 50 sec), annealing (at 65°C for 1 min) and extension (at 72°C for 30 sec), followed by final heating (72°C for 10 min). Amplified libraries were size selected on a 2% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen). Libraries were quantified on a Qubit fluorometer and checked for correct size distribution on a Bioanalyzer. Samples were diluted to a concentration of 4 nM, denatured in 1 N NaOH and 16 pM library, with 5% PhiX control (Illumina, #FC-110–3001) spiked in, were loaded on a MiSeq machine (Illumina, #SY-410–1003) equipped with MiSeq Reagent Kit v3 (Illumina, #MS-102–3001) sequencing chemistry.

Chromatin immunoprecipitation (ChIP)

One mouse hippocampus per ChIP reaction was dounced in 400 μL MNase digestion/ nuclei permeabilization buffer and digested with MNase (Sigma Aldrich, #N3755) to obtain mononucleosomal DNA (~150 bps). ChIP was performed with 4 μL anti H3K9ac polyclonal rabbit antibody (Millipore, #ABE18). DNA from input was extracted in parallel.
A small aliquot of the input samples was checked on agarose gels for size distribution of sheared chromatin.  

**ChIP qPCR**

For quantification by qPCR H3K9ac ChIP DNA was diluted 1:5 with elution buffer (Qiagen) and amplified on a Light Cycler 2.0 (Roche diagnostics, #03531414001) using Quantifast SYBR Green Kit Master Mix (Qiagen, #204054). The following primers were used: *Meis2* (NC_000068.7), positive control locus (present under all sequencing depth settings), fwd 5’-TCGGTCAATATGCCTGTTG-3’, rev 5’-CTGCCCATGCTTGTGTTTC-3’, target locus (present only in the deepest sequencing condition), fwd 5’-GGGCTCTTTAGATGGCAGACT-3’ rev 5’-CAAAATGAATGGGGGTTGGGGG-3’, negative control locus on gene body (present in none of the sequencing conditions), fwd 5’-AAATGTCAACCAGGGACACC-3’, rev 5’-AACCTTTGCAGGCTGGAGTT-3’ and a negative control in an intergenic locus (present in none of the sequencing conditions), fwd 5’-GGGGTGACCGTT-3’, rev 5’-TCGGTCAATATGCCTGTTG-3’, positive control locus, fwd 5’-GGGGTGACCGTTGATCAGTT-3’, rev 5’-CCTATCTCGTCAACCACTGGC-3’, target locus, fwd 5’-GGTTGCTGGAAGGTGGAACGT-3’, rev 5’-TCCCGCATTTGGGCACTATT-3’, locus on gene body, fwd 5’-CGCACTCGTATGAACCT-3’, rev 5’-ATGGTTGGGTGGAATGTCTC-3’, and intergenic locus, fwd 5’-ATTGGTTGGTCAGACCGACT-3’, rev 5’-TACCTTGACCCAGGATGTTG-3’. PCR reactions (10 µL final volume) were run in duplicates with of 1.5 µM of specific primers and 2 µL of ChIP-DNA or input DNA, respectively. ChIP Cts for each sample and primer were normalized by the Ct for the input DNA. Amplification levels are presented relative to the levels for the positive control locus which was set to 1.

**Bioinformatic analysis**

Read quality was checked using FastQC tool.  
Adapters were trimmed using cutadapt.  
SOLiD reads were processed with in house scripts for compatibility with the aligner. The alignment to mm9 and the microbial mix genomes was done using either BFAST v0.7.0a for color space alignment, or BWA.aln v0.7.1022 for nucleotide space with standard parameters. Only uniquely mapping reads were accepted. To estimate fragment size in each library, we used MaSC.  
Reads were elongated to the estimated size using BEADS.  
To calculate the GC distribution, BEADS was used again. Power calculation for H3K9ac was performed with ChIP-Seq Statistical Power (CSSP) analysis in R using the standard settings with two- and fourfold enrichment as parameters.

**Construction of the theoretical GC distribution**

For the construction of the theoretical GC distribution of the microbial genome pool, we used the reference genomes (M. luteus NCCT2665, P. putida KT2440, E. coli K12-DH10B, P. torridus DSM 9790 and P. falciparum Pf3D7_v2.1.5), shredded them to 75 bp to match our sequencing reads using BEADS, and aligned them against themselves to account for mappability issues. The resulting alignment was elongated to 150 bp (to match the shearing of our library) and the GC distribution was calculated again using BEADS.

The estimation of the shift between the theoretical and actually observed GC distribution is made possible by excluding any amplification procedures in our library preparation (before the emulsion PCR). In this type of library, each fragment of the original gDNA mix can either be sequenced once or fail to cluster and drop out. Therefore, we know that the resulting GC distribution after sequencing will not surmount the theoretical distribution at any GC level and we can fit it within using the loess function in R.  
To calculate the shift between the fitted observed GC distribution and the theoretical distribution, we subtracted the areas under the curve (\(\text{AUC}_{\text{theoretical}} - \text{AUC}_{\text{observed}}\)) using R.

This shift between the 2 distributions is used as a proxy for the magnitude of the bias in the analysis.

**Accession numbers**

gDNA and ChIP-seq data have been deposited in the Sequence Read Archive (SRA) under the BioProject accession number PRJNA380045.

**Disclosure of potential conflicts of interest**

No potential conflicts of interest were disclosed.

**Acknowledgment**

We thank Dr. Angel Angelov (Technical University, Munich) for kindly providing us with Pictrophilus torridus.
gDNA, Prof Dr Marc Bramkamp and Dr Karin Schubert (Ludwig-Maximilians-University, Munich) who kindly supplied us with Micrococcus luteus and Pseudomonas putida pellets. Likewise, we thank Dr Berens-Riha (Ludwig-Maximilian-University, Munich) for providing Plasmodium falciparum gDNA. We are grateful to Dr Tobias Spielmann and Florian Kruse (Bernhard Nocht Institute, Hamburg) for kindly preparing large amounts of Plasmodium falciparum gDNA for us. Also, we thank Dr Lutz Wiehlmann (Hannover Medical School, Hannover) for technical support.

Funding

This work was supported by a Marie Curie International Incoming Fellowship within the 7th European Community Framework Program from the European Commission under Grant #332297 (to MJ) and in part by a NARSAD Young Investigator Grant from the Brain and Behavior Research Foundation under Grant #22809 (to MJ). Dr Mira Jakovcevski is an “Attias Family Foundation Investigator.”

References

[1] Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell 2013; 155 (1):27-38; PMID:24074859; https://doi.org/10.1016/j.cell.2013.09.006

[2] PsychENCODE Consortium, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, et al. The Psychencode Project. Nat Neurosci 2015; 18(12):1707-12; PMID:26605881; https://doi.org/10.1038/nn.4156

[3] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011; 12(2):R18; PMID:21338519; https://doi.org/10.1186/gb-2011-12-2-r18

[4] Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One 2013; 8(6):e66621; PMID:23776689; https://doi.org/10.1371/journal.pone.0066621

[5] Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol 2013; 14 (5):R51; PMID:23718773; https://doi.org/10.1186/gb-2013-14-5-r51

[6] Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. Nat Methods 2009; 6(4):291-5; PMID:19287394; https://doi.org/10.1038/nmeth.1311

[7] Huptas C, Scherer S, Wenning M. Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. BMC Res Notes 2016; 9:269; PMID:27176120; https://doi.org/10.1186/s13104-016-2072-9

[8] Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res 2015; 43(21):e143; PMID:26187991

[9] Sims D, Sudbery I, Illot NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 2014; 15(2):121-32; PMID:24434847; https://doi.org/10.1038/nrg3642

[10] Jakovcevski M, Akbarian S, Di Benedetto B. Pharmacological modulation of astrocytes and the role of cell type-specific histone modifications for the treatment of mood disorders. Curr Opin Pharmacol 2016; 26:61-6; PMID:26515273; https://doi.org/10.1016/j.coph.2015.10.002

[11] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D’Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010; 466(7303):253-7; PMID:20613842; https://doi.org/10.1038/nature09165

[12] Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev 2005; 19(5):542-52; PMID:15706033; https://doi.org/10.1101/gad.1272505

[13] Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genom 2012; 13:1; https://doi.org/10.1186/1471-2164-13-1

[14] Hezroni H, Tzchori I, Davidi A, Mattout A, Biran A, Nissim-Rafinia M, Westphal H, Mesheror E. H3K9 histone acetylation predicts pluripotency and reprogramming capacity of ES cells. Nucleus 2011; 2(4):300-9; PMID:21941115; https://doi.org/10.4161/nucl.2.4.16767

[15] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform 2016; 2016:1-12

[16] Reddy AS, O’Brien D, Pisat N, Weichselbaum CT, Sakers K, Lisci M, Dalal JS, Dougherty JD. A Comprehensive analysis of cell Type-Specific nuclear RNA from neurons and glia of the brain. Biol Psychiatry 2017; 81(3):252-64; PMID:27113499; https://doi.org/10.1016/j.biopsych.2016.02.021

[17] Castellanos-Rizaldos E, Milbury CA, Makrigiorgos GM. Enrichment of mutations in multiple DNA sequencing and analysis of factors in sequencing de novo assembly. BMC Res Notes 2016; 9:269; PMID:27176120; https://doi.org/10.1186/s13104-016-2072-9

[18] Jakovcevski M, Ruan H, Shen EY, Dincer A, Javidfar B, Ma Q, Peter CJ, Cheung I, Mitchell AC, Jiang Y, et al. Neuronal Kmt2a/Mll1 histone methyltransferase is essential for
prefrontal synaptic plasticity and working memory. J Neurosci 2015; 35(13):5097-108; PMID:25834037; https://doi.org/10.1523/JNEUROSCI.3004-14.2015

[19] Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[20] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB Net J 2011; https://doi.org/10.14806/ej.17.1.200

[21] Homer N, Merriman B, Nelson SF. BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLoS One 2009; 4(11):e7767; PMID:19907642; https://doi.org/10.1371/journal.pone.0007767

[22] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 2009; 25(14):1754-60; PMID:19451168; https://doi.org/10.1093/bioinformatics/btp324

[23] Ramachandran P, Palidwor GA, Porter CJ, Perkins TJ. MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data. Bioinformatics 2013; 29(4): 444-50; PMID:23300135; https://doi.org/10.1093/bioinformatics/btt001

[24] Cheung MS, Down TA, Latorre I, Ahringer J. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Res 2011; 39(15):e103; PMID:21646344; https://doi.org/10.1093/nar/gkr425

[25] Zuo C, Keleş S. A statistical framework for power calculations in ChIP-seq experiments. Bioinformatics 2014; 30 (6):753-60; PMID:23665773; https://doi.org/10.1093/bioinformatics/btt200

[26] Statistical Research. Random Statistics and Data Science. That's Smooth; 2013 Oct 10 [assessed 2016 Jan 05]. https://statistical-research.com/index.php/2013/10/10/thats-smooth/