A comparison between SOLiD5500XL and Ion Torrent PGM sequencing platforms for the analysis of miRNA expression profiles

Gabriela P. Branco[1]*, Renan Valieris[2], Israel Tojal da Silva[1,2], Jorge?, Gustavo R. Fernandes[1], et al. Maria Galli de Amorim[1], Diana Noronha Nunes[1], Emmanuel Dias-Neto[1,3*]

1 - Laboratory of Medical Genomics, Centro Internacional de Pesquisas, AC Camargo Cancer Center. São Paulo, SP, Brazil.

2 - Laboratory of Computational Biology, Centro Internacional de Pesquisas, AC Camargo Cancer Center. São Paulo, SP, Brazil.

2 – Laboratório de Neurociências Alzira Denise Hertzog Silva (LIM-27). Departamento e Instituto de Psiquiatria, Faculdade de Medicina da Universidade de São Paulo. São Paulo, SP, Brazil.

Equal contribution.

*Corresponding author at the Laboratory of Medical Genomics, Centro Internacional de Pesquisas, AC Camargo Cancer Center. São Paulo, SP, Brazil. Rua Tagua 440, Liberdade. São Paulo, 01508-010, SP, Brazil. Email: emmanuel@cipe.accamargo.org.br

Abstract

Next-generation sequencing (NGS) has certainly been one of the most transformative tools in the field of biomedical sciences. The high-throughput capabilities given by the different NGS platforms, which enabled the sequencing of hundreds of millions of molecules in a single sequencing run, has revolutionized the field of genomics and bioinformatics. In particular, miRNA studies were positively impacted by the ability of gene expression quantification in unprecedented scale without the limitations of closed-platforms. In the present paper, we compare the performance of two NGS-platforms developed by Life Technologies - SOLiD (Sequencing by Oligonucleotide Ligation and Detection) and Ion Torrent PGM (Personal Genome Machine) – for the identification and quantification of microRNAs expressed by two breast-derived cell lines (Hb4 and C5.2). High expression correlation (R^2>0.9) was achieved between SOLiD and PGM for both cell lines, with good expression correlations observed for 97% of the total miRNAs identified by one or another platform. Most of the few discrepancies identified in miRNA expression levels were attributable to low expressed molecules, i.e. those below 3.64 reads/million reads. Quantification divergence indicative of artefactual representation was seen for 14 miRNAs more prevalent in SOLiD and 10 miRNAs more prevalent in PGM. An evaluation of these miRNAs sets regarding nucleotide composition, formation of hairpins or xxxx, revealed no particular features that could explain the better representation by one or another platform. Whereas we highlight the miRNAs that showed quantification discrepancy, we also conclude that the results from both platforms are overall consistent and can be used interchangeably in microRNA-expression studies.

1.Introduction

Over the last years the scientific community has produced a remarkable amount of genomic information that helped the understanding of many fundamental medical questions and biological phenomena. This has been made possible by the development of new genomic technologies, which led to dramatic cost-reduction and allowed the deep exploitation of genomes, exomes and transcriptomes in all areas of biological research (van Dijk et al., 2014).

However, the massive amount of data produced by the so-called Next-Generation-Sequencing (NGS) platforms also brought significant challenges regarding data-storage, analysis and database management solutions. Furthermore, the use diverse library construction protocols and sequencing chemistry approaches required by the distinct NGS platforms, resulted in a vast amount of data characterized by high variability in terms of read-length, error rates, possible representation biases and different error profiles (Shendure & Ji, 2008; Quail et al., 2012; Yang et al., 2013; PMID: 23592973; PMID: 22522955; PMID: 23405114; PMID:22726842). This brings an important challenge for the interchangeable use and sequencing platform comparison for the assessment-power of publically available data.

Whereas specific databases have been created for the public availability of these reads (e.g. the Sequence Read Archive; www.ncbi.nlm.nih.gov/sra), allowing the free-use of the data, studies directed to a systematic comparison of data derived from different platforms are needed to point platform-dependent discrepancy and to determine how equivalent are the data produced by different sequencing strategies.

Although papers have compared miRNA detection and quantification using platforms such as NGS, microarrays and nCounter (Nanostring) (Chatterjee et al., 2015; Nassirpour et al., 2014; Willenbrock et al., 2009) few manuscripts have systematically compared the sequencing of the same source samples by distinct NGS platforms. Here we compare micro-RNA (miRNA) sequencing data, generated from the same two breast cell lines (HB4 and C5.2), after large-scale sequencing with two NGS platforms: SOLiD (Sequencing by Oligonucleotide Ligation and

Detection) and Ion Torrent PGM (Personal Genome Machine), both produced by Thermo (*Thermo Fisher Scientific, USA*), and delve into the analysis of how comparable are the data produced by both platforms. We should note that whereas the Ion Torrent PGM platform is currently in use by many institutions in the world, the SOLiD platform has been discontinued. However, the analyses performed here are still valid as there are xxxx studies and xxx million SOLiD-derived sequences deposited in public databases (name them?), and this number has increased recently after the completion of some analyses and the deposition of the data.

2. Materials and Methods

2.1 Cells

The study was performed with the mammary cell lines HB4a and C5.2. C5.2 is a cell clone derived from the transfection of mammary epithelial origin HB4a with ERBB2/HER-2 oncogene (**HARRIS** *et al*, 1999). Cells were grown at 37ºC and 5% CO_2 in RPMI 1640 medium supplemented with 10% fetal bovine serum, 1% antibiotic-antimycotic (penicillin/streptomycin/amphotericin-B; Invitrogen, Carlsbad, CA, USA) and 5mg/ml hydrocortisone (Sigma-Aldrich, St. Louis, MO, USA) (Carraro *et al*, 2010).

2.2 RNA extraction and quantification

miRNAs were extracted using the miRNeasy mini kit together with the RNeasy MinElute cleanup kit in the QIAcube equipment (Qiagen, Hilden, Germany), following the provided instructions. miRNA quantifications were performed using the 2100 bioanalyzer small RNA chip (Agilent, Santa Clara, USA). Aliquots of 100ng of small RNAs, derived from the same RNA-extraction procedure, were used for the simultaneous preparation of miRNA libraries as follows.

2.3. PGM Ion Torrent - miRNA libraries construction and sequencing

For library construction, we have followed the recommendations of the Ion Total RNA-Seq Kit (Life Technologies, Carlsbad, California, USA). Ion OneTouch 200 Template Kit v2 DL was used for emulsion PCR and sequencing was performed with Ion PGM 200 Sequencing Kit (180 flows).

2.4 SOLID 5500xl - miRNA libraries construction and sequencing

The SOLiD Seq Total RNA Kit (Life Technologies, Carlsbad, CA, USA) was used to prepare the miRNA libraries. Some modifications were implemented in the protocol, as follows: A) Hybridization and RNA binding: miRNAs contained in ≤1µl of enriched small RNAs samples were hybridized and ligated to adapters. For each reaction, we used 3µl of hybridization solution, plus 2µl of the SOLiD™ adaptor mix and water to a final volume of 8 µl. The reaction volume was incubated at 65°C for 10 min and transferred directly to the ice. Subsequently we added 10 µl of 2X ligation buffer and 2 µl of ligation enzyme mix to each reaction, followed by incubation at 16°C for 16 h. B) Reverse transcription contained: 4 µl of reverse transcription buffer 10X; 2µl of dNTP mix (2.5 mM); 2 µl of reverse transcription primer SOLiD™ and 11µl of nuclease-free water. After incubation at 70°C for 5 min, we added the 1µl of the ArrayScript™ Reverse Transcriptase and incubated for 30 min at 42°C. C) cDNA purification, size selection and amplification: cDNAs synthesized in the previous step were column-purified with MinElute PCR purification kit (Qiagen, Hilden, Germany). For size selection, 5 µl of the cDNAs were combined with 5 µl of sample buffer (2X Novex® TBE – urea sample buffer), the mixture was heated (95°C for 3 minutes) and immediately transferred to ice. Samples were fractionated using the XCell SureLock™ system mini-cell with polyacrylamide gels – (10% Novex® TBE Urea Gel 1.0 mM, 10 well in Novex® TBE running buffer for 1 h at 180V. We subsequently stained the gels in the same running buffer (1X) containing 5 µl of SYBR® Gold nucleic acid gel stain for 10 min. Bands were visualized with the safe blue-light imager transilluminator (all from Invitrogen, Carlsbad, CA, USA) and cDNA fragments ranging from 60 to 70nt -- corresponding to miRNAs ligated to adapters -- were excised and amplified as recommended. Amplicons of two independent PCRs were combined and mixed with 1.8X volumes of the Agencourt® AMPure® XP Beads (Beckman Coulter, Brea, CA, USA) and incubated for 5 mins at room temperature. The beads containing amplicons were washed with ethanol, and unbound products were purified again with the same beads (ratio 2:1). Products of interest were eluted in 20µl of 1X low TE and evaluated with the high-sensitivity bioanalyzer chip, as recommended in the protocol. D) The E20 emulsion PCR (ePCR) and ePCR enrichment were performed following the recommended protocols (Applied Biosystems, USA).

Sequencing was performed according to the protocol 5500 Series Genetic Analysis System User Guide (Applied Biosystems, USA).

2.4 Bioinformatics and statistical analyses

miRNAs were identified from both SOLiD- and PGM-derived reads after quality-filtering, adapter removal and mapping against miRBase (release 20) using miRDeep2 (Friedländer et al. 2012; version 2.0.0.5). Default parameters were used allowing a maximum of one mismatch, a seed sequence of 18 nt and no mismatches on seed. Aligned reads were filtered to capture mature miRNAs preferentially represented by one or another platform in one of the cell lines as well as unique-to-platform miRNAs differentially represented by one of the platforms consistent for both cell lines. Due to intrinsic differences of PGM and SOLiD, which have very distinct throughputs (SOLiD yielded 4 times more sequences than PGM), we have applied stringent requirements for determining the presence of specific miRNAs in this dataset. Therefore, in order to compare the miRNA represented by each NGS platform, a miRNA was considered if at least two miRNA-corresponding reads were available from PGM-data and at least 12 reads were derived from SOLiD (4x difference - throughput correction coefficient the considers the sequence throughput achieved here for the different data sets).

To avoid the mapping of sequencing reads to multiple miRNAs that differ from each other by a single nucleotide, the mature sequences of all miRBase miRNAs were clustered, allowing a maximum of 1 nt mismatch, and provided a list of 40 distinct miRNA clusters that have a maximum distance of 1nt (Supplementary Table X) as well as 2692 individual miRNAs (Supplementary Table 1). Read counts were calculated for each miRNA (clustered or unique). After sequence alignment and annotation, R (v2.12) and perl (v5.14.2) scripts were applied to standardize miRNA nomenclature and counts across all samples.

Saturation/rarefaction curves were built using all mature miRNAs identified, with no limits for a minimum of reads. We employed an R script to sample 100 points, 100 times each. The sampling method was done with replacement and the relative frequency of each miRNA was used as its sampling probability. Correlation curves were drawn using log2 transformation of the number of reads normalized per million-reads sequenced in each sample. Outliers were considered when the

normalized count difference between samples was greater than the 1.5*IQR + Q_3 value, where IQR is the interquartile range (Q_3-Q_1) and Q_1 and Q_3 are the first and third quartiles. The inner diagonal line represents the line y=x and each line represent the outlier threshold.

ANOVA tests were used to evaluate possible associations between physico-chemical parameters of the outlier miRNAs observed for the distinct sequencing platforms. The parameters evaluated were: i) GC% content; ii) the presence, composition and length of homopolymers; iii) free energy; iv) formation of loops….. The Fisher's Exact Test was used to evaluate differences which parameters had most effects over their representation by the Ion PGM or SOLiD sequencing platforms. For all tests, p-values <0.05 were considered to be significant.

3. Results

After quality and size filtering we evaluated 7,883,393 reads provided by SOLiD and 1,924,046 reads provided by the Ion PGM. The percentage of unmapped reads or sequencing reads removed due to low quality, for all cells and platforms are shown in **Supl. Figure 1**. **Table 1** shows the number of reads and the set of expressed miRNAs determined by miRDeep2 for both cell lines by these two sequencing platforms, including the number of distinct miRNAs identified by both platforms for both cell lines, as well platform-specific miRNAs. Renan, vc pode verificar se isto está correto para p mínimo de 2 reads?

Supl Figure 1 – Distribution of filtered reads from PGM and SOLiD platforms for both cell lines

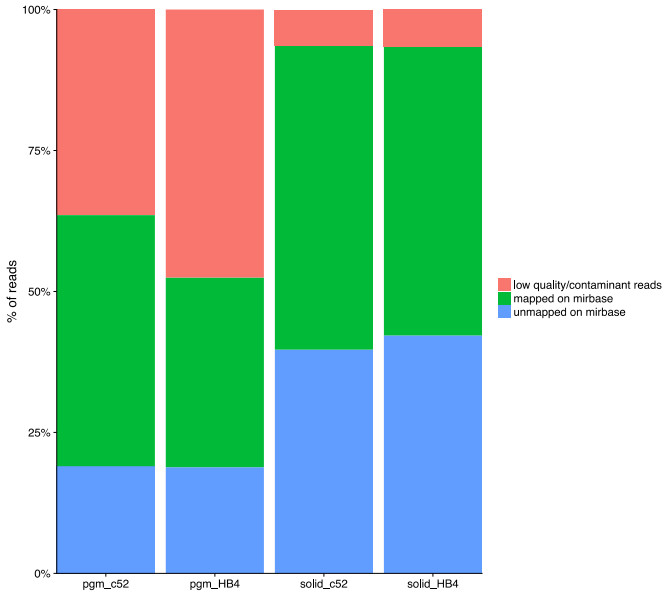


Table 1 – Sequencing results and miRNA identification in Hb4a and C5.2 cell lines using PGM and SOLiD NGS platforms.

	HB4a	C5.2 (%)	
Valid reads (PGM)	1,099,181	824,865	
Valid reads (SOLiD)	3,501,788	4,381,605	
Ratio SOLiD/PGM reads[1]	3.19	5.31	
Distinct miRNAs identified by PGM[2]	416	429	
Distinct miRNAs identified by SOLiD[3]	407	438	
Total distinct miRNAs (SOLiD + PGM)	465	495	
miRNAs more abundant in PGM[4]	3 (0.6%)	3 (0.6%)	4 (0.8%)
miRNAs more abundant in SOLiD[5]	3 (0.6%)	6 (1.3%)	5 (1.1%)

[1]Used to adjust the corresponding number of reads which is discrepant due to the distinct throughputs of the two platforms. [2]Considers only miRNAs identified by at least 2 reads. [3]Considers only miRNAs with read counts equal or above xx, after applying the throughput correction coefficient. [4]Considering a total of 491 distinct miRNAs identified using PGM. [5]Considering a total of 478 distinct miRNAs identified using SOLiD.

Besides the inherent throughput from the Ion-PGM and the SOLiD NGS platforms, we observe from the saturation plots shown in **Figure 1** that a very similar

coverage trend and saturation profiles were reached; with very similar trends to point the number of expressed miRNAs. The saturation plots from each platform were also consistent to indicate the higher number of miRNAs expressed in the HB4a cell compared to C5.2 (**Figure 1**), even with the different converage given bu SOLiD and PGM.

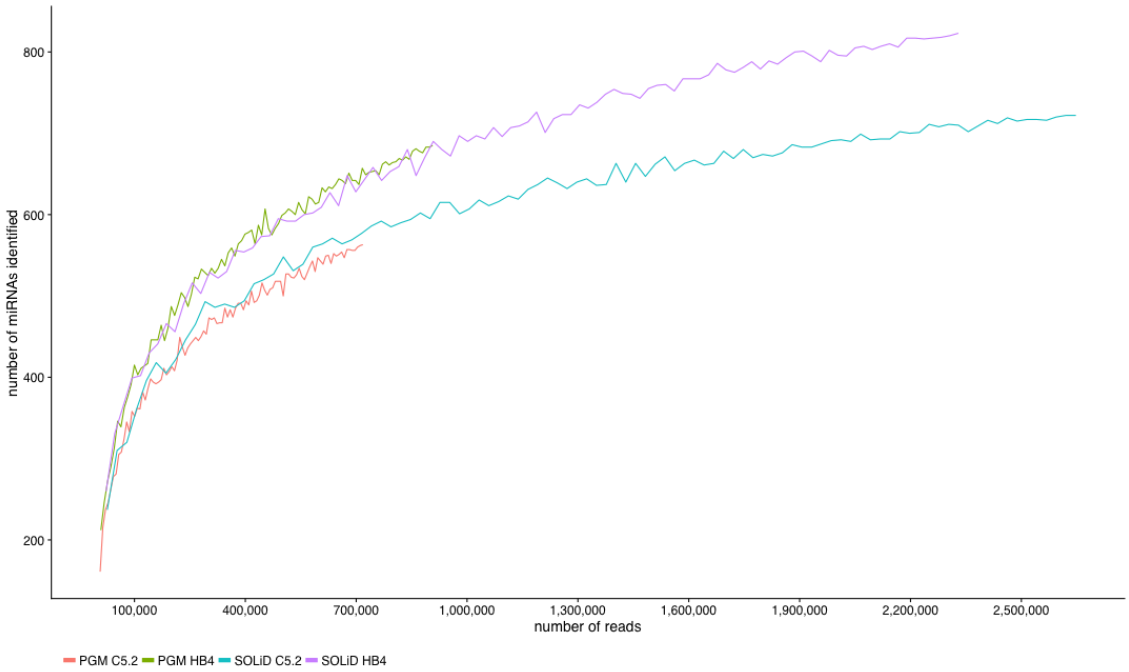
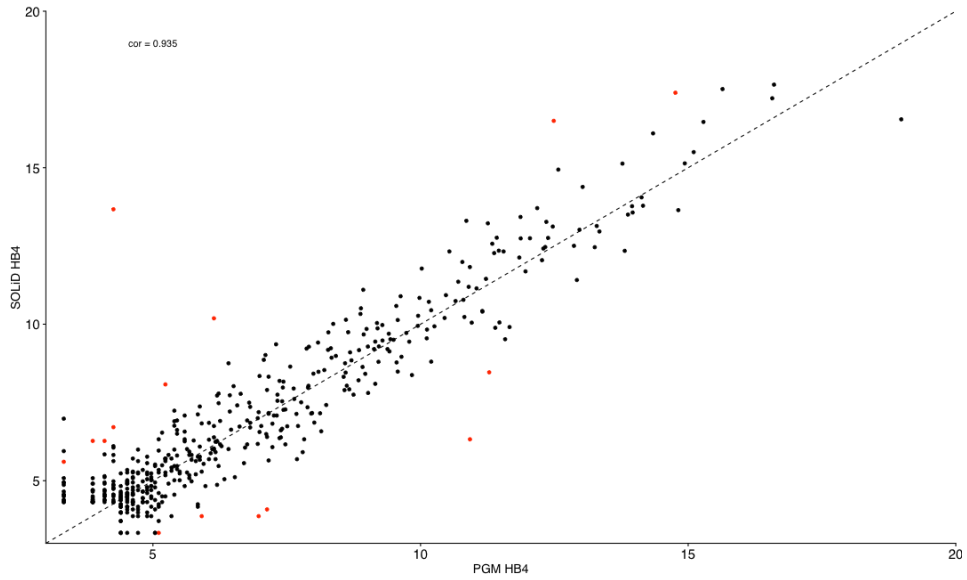


Figure 1 – Saturation plots of miRNAs identified by both sequencing platforms using Hb4 and C5.2 cell lines considering all sequencing reads mapped to mature miRNAs from mirBase.

Our analyses demonstrate that both sequencing platforms allow a robust representation of miRNAs by in terms of number and abundance of the identified miRNAs. This can also be seen in the correlation plots (**Figure 2**) and Venn-diagrams (**Figure 3**) that indicate very few miRNAs to be discrepant quantitatively between these platforms.

a)



b)

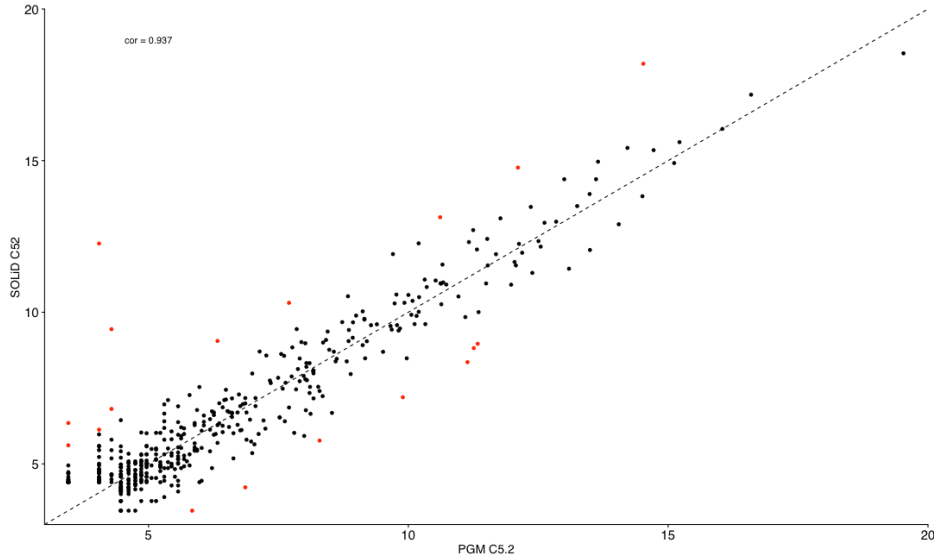


Figure 2 – Plots showing the miRNA expression-correlation analysis between SOLiD and PGM platforms for the cell lines Hb4 (2a) and C5.2 (2b). Expression levels for each miRNA are given in reads/million. miRNAs with differential representation in each platform (DeSeq) are indicated in red. The Pearson correlation coefficient (cor) was calculated tor each cell line.

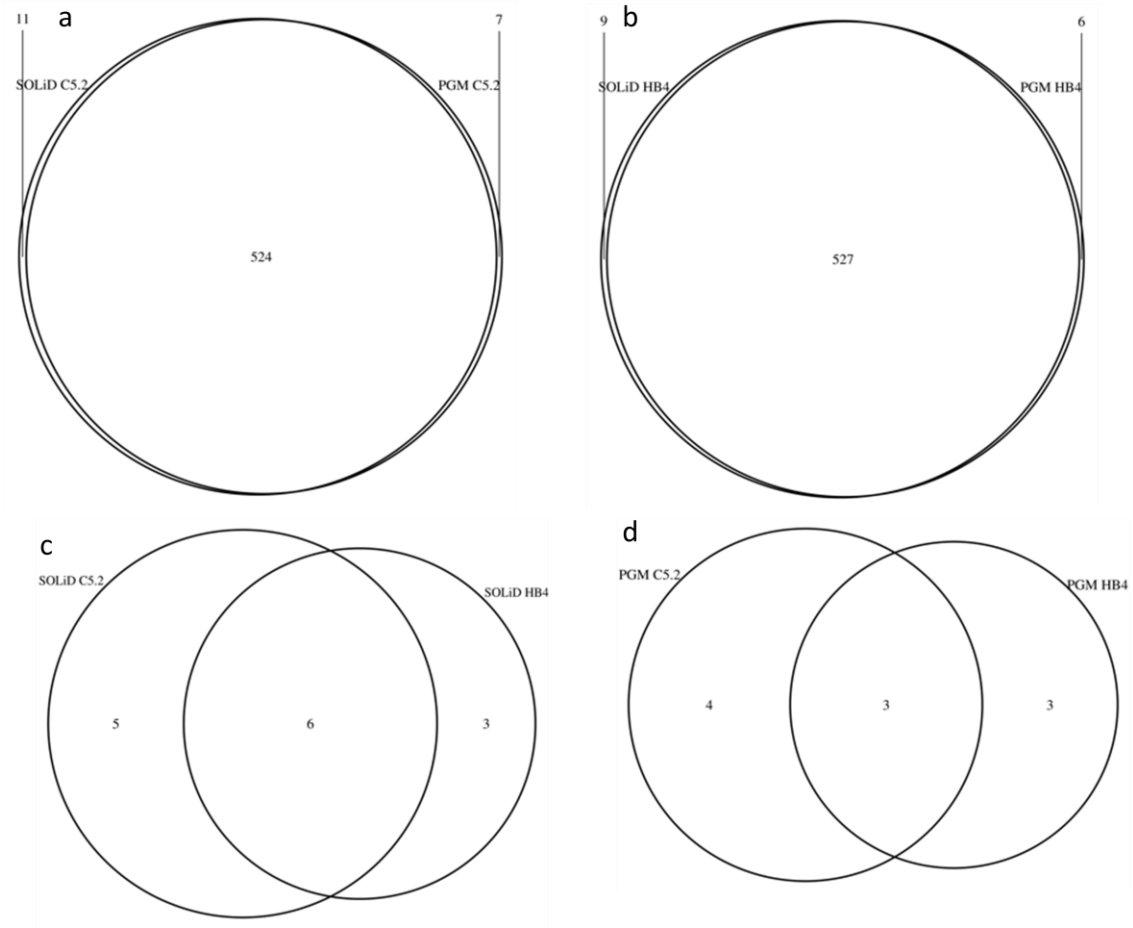


Figure 3 – Venn diagrams depicting the set of miRNAs preferentially represented in C5.2 (a) or Hb4a (b) cell lines using SOLiD or PGM platforms. miRNAs more abundantly represented in SOLiD (c) or PGM (d). The miRNAs indicated here correspond to those indicated by red dots in Figure 2.

Our results are in agreement with the expected similarity of Hb4a and its derived clone C5.2, which differ only in respect to the overexpression of *ERBB2* in the later, and show a very similar miRNA expression profile for both cells. Using the criteria adopted here, we found exactly the same number of expressed miRNAs (542 – **Figure 3a,b**) for both cell lines. Out of these 542 miRNAs, 2.76% and 3.3% were differentially represented by the NGS platforms (respectively in Hb4a and C5.2). The full list of miRNAs identified for both cell lines, as well as their normalized level of expression for both platforms is given in **Supl. Table 1**. The list of miRNAs

with detectable differences in representation by one or another NSG platform is given in **Tables 2 and 3**.

Table 2 – miRNAs more abundantly detected by the SOLiD platform

miRNA (hsa-miR)	Fold change (SOLiD:PGM)	p-value	Reads/million (SOLiD)	Reads/million (PGM)	Cell line	Sequence
103a-3p	8.41x	0.0060	128212	15233	C5.2	agcagcauuguacagggcuauga
	4.03x	0.0406	70822	17552	Hb4a	
107	3.82x	0.0324	3831.24	1000.16	C5.2	agcagcauuguacagggcuauca
1249	16.13x	0.0180	39.03	2.42	C5.2	acgcccuucccccccuucuuca
142-5p	N.A.	0.0302	12.57	0	Hb4a	cauaaaguagaaagcacuacu
150-5p	N.A.	0.0084	26.02	0	C5.2	ucucccaacccuuguaccagug
	26.36x	0.0028	23.99	0.91	Hb4a	
152-5p	8.07x	0.0202	102.8	12.74	Hb4a	agguucugugauacacuccgacu
181b-3p	13.18x	0.0424	23.99	1.82	Hb4a	cucacugaacaaugaaugcaa
199a-3p/ 109b-3p	17.6x	0.0429	21.23	1.21	C5.2	acaguagucugcacauugguua
	12.86x	0.0274	35.12	2.73	Hb4a	
223-3p	N.A.	0.0340	12.55	0	C5.2	ugucaguuugucaaauacccca
29a-3p/ 29c-3p	4.19x	0.0266	11927	2849	C5.2	uagcaccaucugaaaucgguua
	10.48x	0.0062	37988	3622	Hb4a	uagcaccauuugaaaucgguua
301b	4.39x	0.0270	532.68	121.23	C5.2	cagugcaaugauauugucaaagc
3607-5p	1731x	0.0367	2094.4	1.21	C5.2	gcaugugaugaagcaaaucagu
	1968x	0.0286	5353.8	2.73	Hb4a	
4284	118.8x	0.0002	287.57	2.42	C5.2	gggcucacaucaccccau
	14.39x	0.0042	471.19	32.75	Hb4a	
4521	5.61x	0.0200	217.73	38.79	C5.2	gcuaaggaaguccugugcucag

Table 3 – miRNAs more abundantly detected by the PGM platform

miRNA (hsa-miR)	Fold change (PGM: SOLiD)	p-value	Reads/million (PGM)	Reads/million (SOLiD)	Cell line	Sequence
let-7i-5p	8.16x	0.0060	1659.7	203.35	C5.2	agcagcauuguacagggcuauga
	11.43x	0.0406	1563.0	136.79	Hb4a	
1295-5p	12.75x	0.0324	189.12	14.83	C5.2	agcagcauuguacagggcuauca
1307-5p	8.59x	0.0180	1574.8	183.27	C5.2	acgcccuucccccccuucuuca
16-1-3p	45.13x	0.0084	61.83	1.37	C5.2	ucucccaacccuuguaccagug
	67.03x	0.0028	76.42	1.14	Hb4a	
200c-3p		0.0202			Hb4a	agguucugugauacacuccgacu
20a-3p		0.0429			C5.2	acaguagucugcacauugguua
		0.0274			Hb4a	
25-5p		0.0340			C5.2	ugucaguuugucaaauacccca
3613-5p		0.0062			C5.2	uguuguacuuuuuuuuuuguuc
424-3p		0.0270			C5.2	cagugcaaugauauugucaaagc
4455		0.0472			Hb4a	agggugugugugu uuuu

is given annotated for each cell line using both NGS platforms, including the overlaps found for the platforms and for the cell lines are indicated in **Figure 1**.

Next we evaluated the expression correlation and differential representation of the miRNAs identified by both platforms using Variance Stabilizing Transformation (VST) of expression, as a means to have precise correlation metrics and to verify if expression levels could affect the differential representation between these sequencing platforms. As shown in **Figures 2A** and **2B**, good expression correlations have been found for the NGS platforms for both cells. The figure also indicates that the majority of miRNAS with differential representation between the platforms have low levels of expression (70% and 72% are below 10 normalized counts, respectively for HB4a and C5.2). However a few very discrepant cases remain such as hsa-miR-3607-5p with 3 and 1 counts per million in PGM versus

5436 and 2110, respectively for HB4a and for C5.2 (see **Tables 2 and 3**, as well as arrows in **Figures 2A and 2B**).

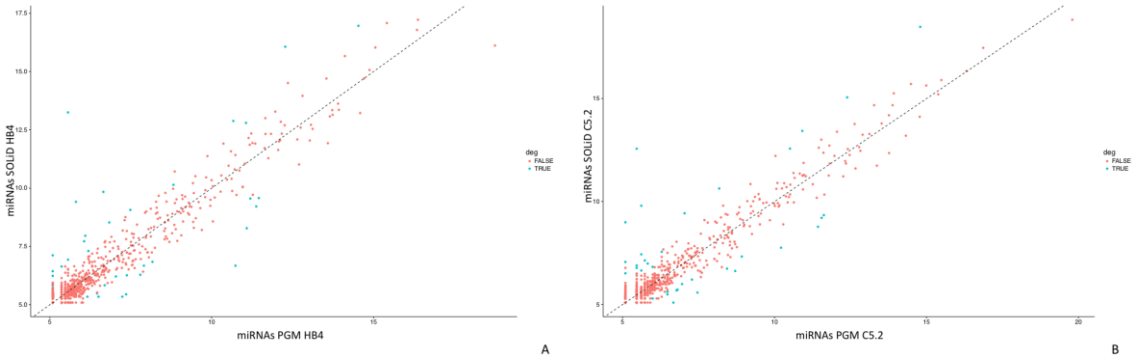


Figure 2 – miRNA expression correlation between PGM and SOLiD sequencing platforms for both cell lines. Expression correlations are shown for HB4a (R^2=0.92) (A) and C5.2 (R^2=0.949) (B) cell lines. The colors of the dots indicate miRNAs with (blue) or without (red) statistically significant expression differences. Add arrow to point specific miRNAs for us to comment in the text.

In order to investigate factors that could affect the differential representation of miRNAs by these platforms, we investigated a number of possible physical-chemical characteristics of the discrepant miRNAs. The frequencies of nucleotides (adenine, guanine, cytosine and uracil) or nucleotide pairs (G+C x A+T) was analyzed for the 26 outlier miRNAs preferentially represented by SOLiD (17 miRNAs) or by PGM (9 miRNAs). We found xxxxxx no statistical differences between the base composition of these miRNAs (**Figure 3**).

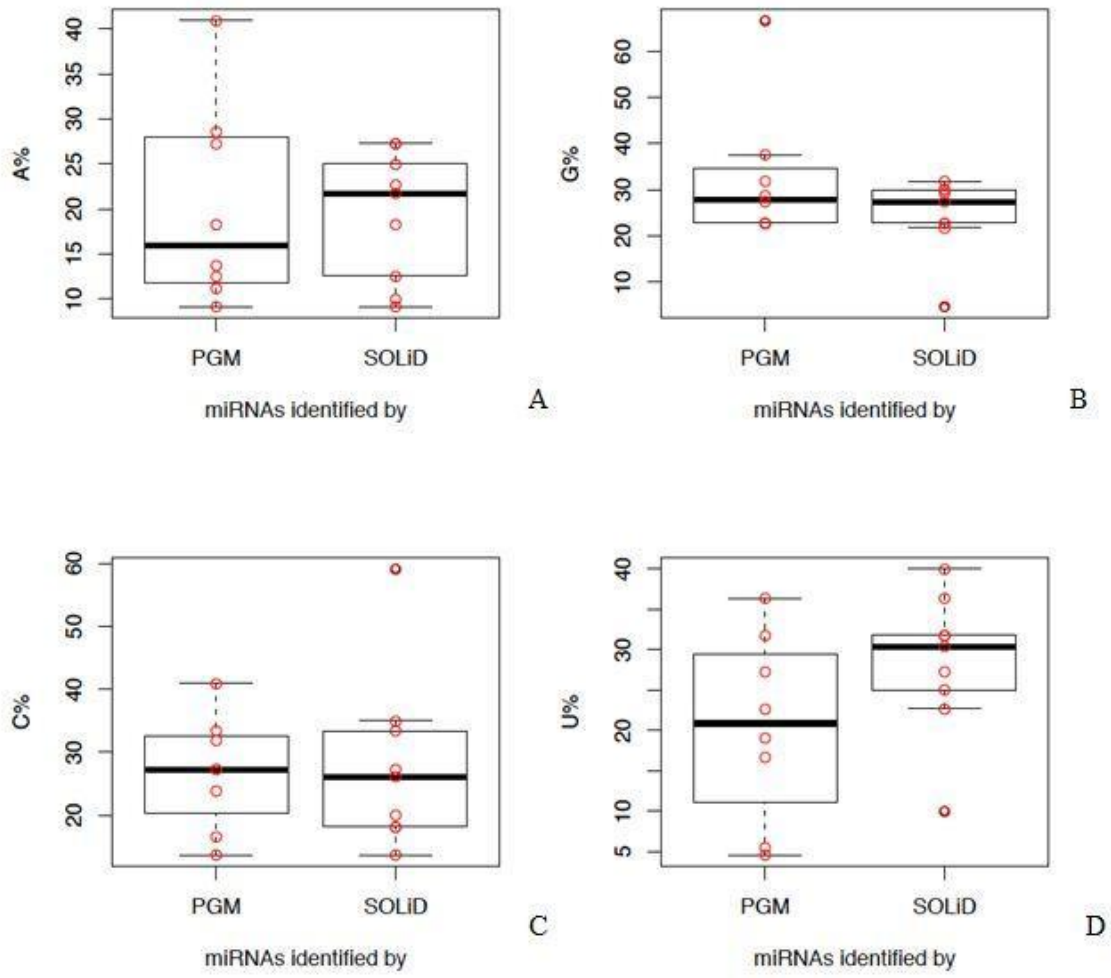


Figure 3 - Comparison nucleotide composition of differentially represented miRNAs identified in both cell lines (HB4a and C5.2) by PGM and SOLiD. A- % adenine p= 0.8566; B- % guanine p= 0.2004; C- %cytosine p= 0.8563; D-%uracil p= 0.1388. Add G+C content repetir com dataset maior

As RNAs are single stranded molecules with a tendency spontaneously twist into secondary/tertiary structures, we evaluated the Gibbs free energy (reference) of the mature miRNAs to evaluate if denaturing conditions of the distinct NGS approaches would favor/disfavor the representation of some miRNAs in all miRNAs identified in HB4a and C5.2 by PGM and SOLiD was also compared, but we found no statistical difference between the two sequencers (p = 0.9856, Figure 4). The same analysis was performed with miRNAs identified in only one cell line (HB4a or C5.2), but it also showed no statistical difference between the two sets (p = 0.7614, Figure 4). Repetir com o set maior de 266 x 46

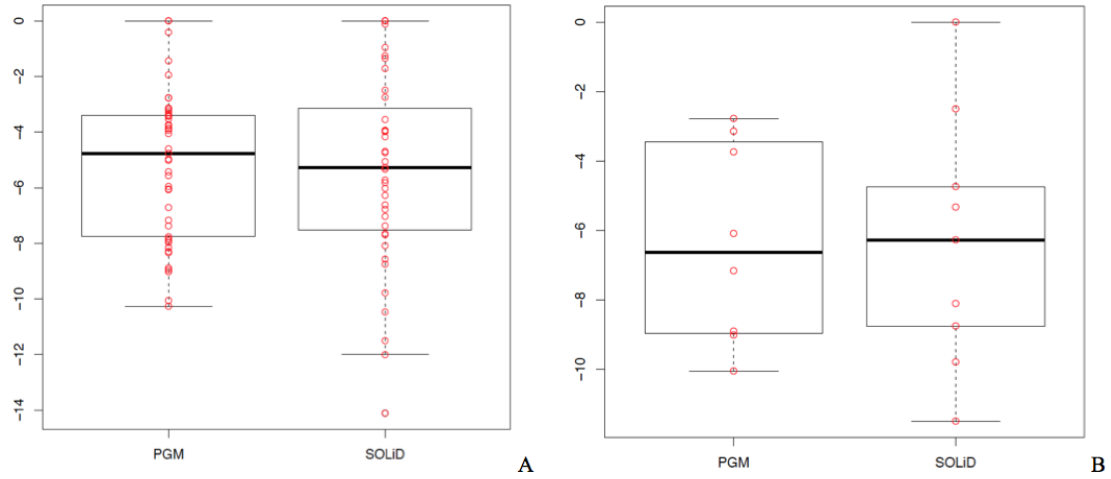


Figure 4 - A - Comparison of Gibbs free energy in miRNAs identified in HB4a or C5.2 by PGM and SOLiD. B - Comparison of Gibbs free energy in all miRNAs identified in HB4a and C5.2 by PGM and SOLiD.

Other features were also compared between differentially represented miRNAs the data obtained with the two sequencers: i) miRNA length, ii) composition, location, number and length of homopolymers and of predicted loops, and the regions in which homopolymers were found. Those comparisons were performed based on the 17 miRNAs outliers identified in both cell lines, and no significant differences were found, with an exception for the number of loops, which differ significantly between sequencers (p=0.0041, Supplementary Tables 1–6). Although re, this difference is probably an artifact caused by the small number of miRNAS used: when these analysis were repeated with all miRNAs outliers identified by the two sequencers in any one of the cell lines (n=84), no significant differences were found in any of the features (Supplementary Tables 7–12). Repetir com dataset maior

Discussion

NGS technologies have strongly impacted genomics and will have far-reaching value in many areas of biological and biomedical research. The high coverage and accuracy now provided by NGS is likely to make the data useful for many years to come. In this sense it is relevant to mention that the analyses of public databases (names here) we revealed above xxx billion reads available, derived from SOLiD xx

of these constitutes miRNA-datasets for diverse conditions.

In the field of transcriptome research, when the aim is to determine differentially expressed transcripts, NGS provides a very appealing approach due to its intrinsic openness where the transcripts to be evaluated are not restricted to those available in closed platforms such as qRT-PCR or hybridization arrays, dynamic range of detection and other benefits.

The chemistries used by ABI SOLiD and Ion PGM platforms have been described elsewhere (refs here). In brief, the ABI SOLiD makes use of a sequencing-by-ligation approach in which an emulsion-PCR (ePCR) with small magnetic beads is used to amplify the DNA fragments for parallel sequencing. During SOLiD sequencing, DNA ligation is carried out to link specific fluorescent labeled 8-mer oligonucleotides for "dinucleotide-encoding", whose 4th and 5th bases are encoded by specific fluorescence. Each fluorescent marker on an 8-mer identifies a two-base combination, which can be further distinguished with a universal primer-offsetting scheme. This allows a universal primer that is offset by one base from the adapter-fragment position to hybridize to DNA templates in five cycle sets permitting the entire fragment to be sequenced and each base position sequenced twice during each cycle. Each ligation step is followed by fluorescence detection and another round of ligation. The chemistry allows read lengths with lengths varying from xxx to xxx and between 80–100 Gbp of mappable sequences are produced per run or over 2 billion reads per run with a raw base accuracy of 99.94% due to its 2-base encoding mechanism. Besides the high sequencing accuracy and the high throughput, this technology is no longer in use as it is extremely laborious and almost a month is required just to perform a sequencing run.

The Ion Personal Genome Machine (PGM) was the first available NGS platform that uses no fluorescence or image capture (LIU et al, 2012; ROTHBERG et al, 2011). The sequencing is based on the use of a semiconductor chip, which detects the reduction of pH when an ion proton is released right after the incorporation of a nucleotide by the polymerase. The advantages of the PGM include it's small size, the capability of producing long reads (>400nt), a higher speed (2-4 hours runs), a good accuracy of 98% and lower cost ($1 per 1 million bases) (LIU et al. 2012;

QUAIL et al., 2012), this platform has grown recently especially for clinical applications, small laboratories and for the investigation of less-diverse genomes or transcriptomes (LIU et al. 2012). Although, PGM presents a significant homopolymer-associated indel errors (1.5 errors per 100 bases) (LOMAN et al., 2012 – update refs here), and this may affect the correct identification of InDels or the mapping of shorter molecules such as miRNAs.

Due to intrinsic advantages given by the use of NGS to study miRNAs -- such as the identification of mutations, polymorphisms, miRNA-editing, expression levels and even the identification of new miRNAs (**HUANG** *et al* 2009.; **CREIGHTON** *et al*. 2009; **HAGER** 2009; **LU** *et al*. 2009; **FRIEDLANDER** *et al.* 2008; **WEI** *et al.* 2009; **MEIRE** *et al*. 2010) -- this approach is rapidly replacing others such as qRT-PCR arrays and microarrays. However, the full transcriptional characterization of miRNAs has been partially limited by the complexity and increased time-requirements of available RNA-seq library construction methods used in NGS, which also seem to have a systematic biased representation of miRNAs (**LESHKOWITZ** et al. 2013; **LINSEN** *et al*. 2009; **TIAN** *et al*. 2010; **HAFNER** *et al*. 2011; **VAN NIEUWERBURGH** *et al*. 2011). These trends may be introduced during PCR amplification, ligation, and cDNA synthesis steps (**LESHKOWITZ** *et al*. 2013).

Rarefaction curves ?? number of reads for cell line HB4a and C5.2. B- rarefaction curves percentage of reads for cell line HB4a and C5.2. C- correlation between PGM and SOLiD data cell line C5.2 ; D- correlation between PGM and SOLiD data cell line HB4a..

Next we generated saturation curves based on data from HB4a and C5.2 cell lines sequenced by both platforms. The curves show that, when the same number of reads was used for both platforms, saturation was reached with XXX (%) reads and indicates the presence of approximately xxx (%) and xxx (%) distinct miRNAs, respectively for HB4a and C5.2 (Figure 1a and 1b).

“The next question to be investigated was whether the concordance in detection was higher for highly expressed miRNAs, since these miRNAs are likely to be detected more easily. A similar overlap analysis was therefore performed between the top 25% of detected miRNAs for each platform. Surprisingly, the percentage overlap between the platforms decreased when investigating the top 25% of

miRNAs detected by each platform".

"This study has shown that high quality data can be generated, along with high confidence in the detection of miRNAs, even with relatively degraded clinical grade FFPE samples. The three platforms, NGS, microarray and NanoString detected a maximum of 345, 125 and 411 miRNAs respectively. There was some variation in the identities of the miRNAs detected in each platform, with a total of 98 miRNAs concordantly detected by all three platforms. Within each platform reproducibility was very high, and across platforms concordance and shared detection remained high between any two of the three platforms, with significant shared detection occurring between each of the platforms. Since the three platforms involve different methodologies, there is significant potential for methodological bias to influence the results. For example, NGS involves an amplification step, whereas NanoString does not, while microarray involves a hybridization step."

Conclusions

Above 9x% of the miRNAs identified for the two cell lines studied here showed the same levels of expression by SOLiD and PGM platforms. For the miRNAs suggested as specific to one or the other cell line xx % were indicated by both platforms. An analysis of miRNAs suggested only by SOLiD or only by PGM revealed no platform-specific trends in terms of GC content, homopolymer content, A, G, C or T content as well as the predicted formation of secondary structures.

XX% of the miRNAs platform-specific miRNAs showed expression levels below xx of expression, suggesting that coverage is more relevant to the identification of miRNA species than the other factors evaluated here.

Competing interests

Authors declare no competing interests.

Authors' contributions

Conception of the study: EDN. Execution of the RNA-seq experiments: GPB, MGA. Data analysis: GPB, GFR, RV, ITS. Manuscript writing and editing: GPB, DNN, EDN. Study supervision: DNN, EDN. All authors have read and approved the final manuscript.

5. Acknowledgements

This project was funded by CNPq, Associação Beneficente Alzira Denise Hertzog Silva (ABADHS), Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES, Brazil - numero do processo da bolsa bolsa Gabi) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - numero do processo da bolsa Gabi). EDN is a research fellow of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

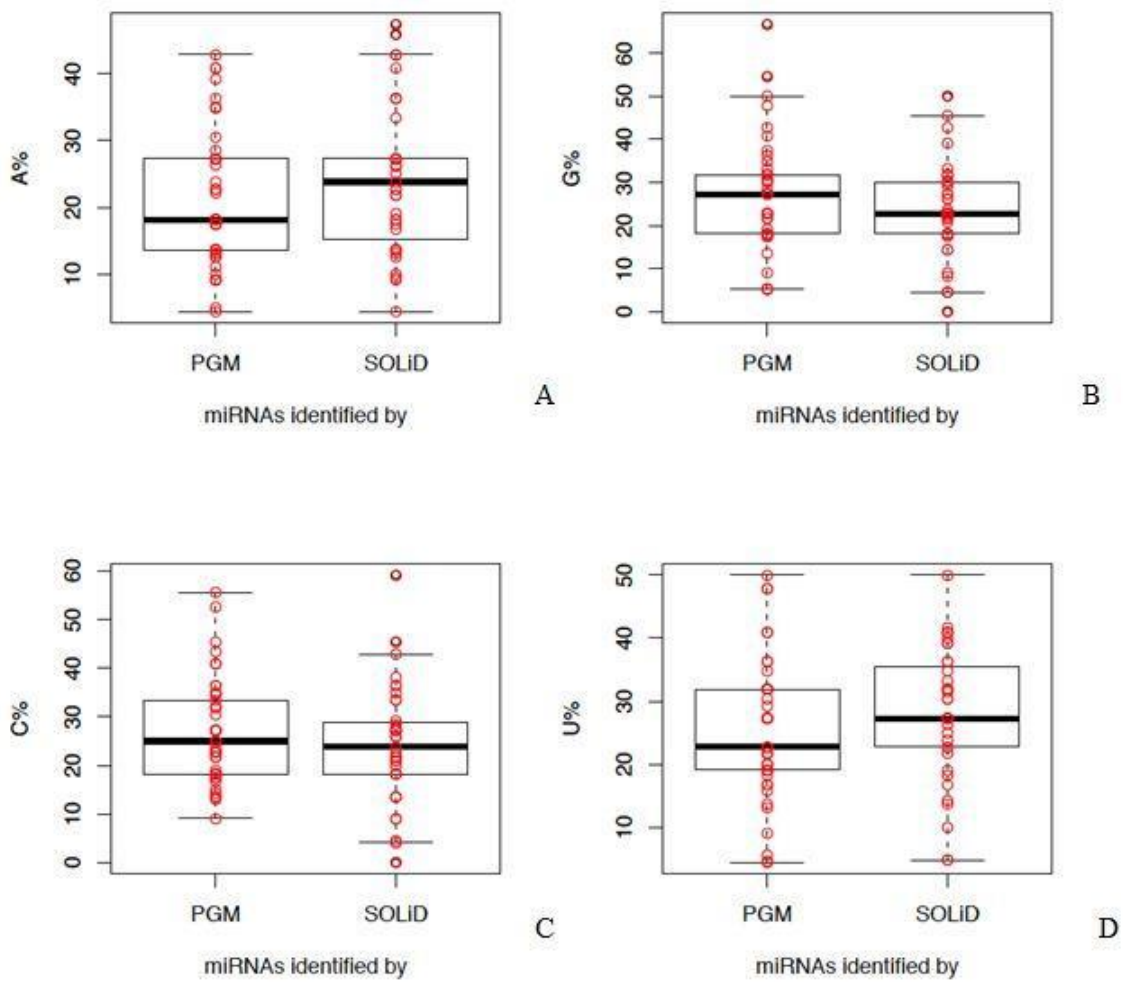


Figure 5 - Comparison of percentage of nucleotides in outlier miRNAs identified by PGM and SOLiD in HB4a or C5.2. A- % adenine p= 0.0789; B- % guanine p=0.186; C- %cytosine p= 0.4393; D- %uracil p= 0.5504.

Supplementary

Supplementary table 1 – Size miRNAs outliers identified in HB4a and C5.2 by SOLiD and PGM, P= 0,7374.

Size miRNA (number of nucleotides)	PGM	SOLiD
18	1	0
20	0	2
21	1	0

22	5	5
23	0	1
24	1	1

Supplementary table 2 – Homopolymer length of differentially represented miRNAs identified in HB4a and C5.2 by SOLiD and PGM, P= 0,3608

Size homopolymer	PGM	SOLiD
–	3	6
4	4	1
5	1	0
6	0	1
7	1	1

Supplementary table 3 – Number of homopolymers in the miRNAs outliers identified in HB4a and C5.2 by SOLiD and PGM, P= 0,4558

Supplementary table 4 – Number of loops in the miRNAs outliers identified in HB4a and C5.2 by SOLiD and PGM, P= 0,004

Loops amount	PGM	SOLiD
0	0	1
1	8	2
2	0	6

Supplementary table 5 – Base of homopolymers in the miRNAs outliers identified in HB4a and C5.2 by SOLiD and PGM, P= 0,3504

Base homopolymer	PGM	SOLiD
–	3	6
a	2	0
c	1	1
g	2	0
u	1	2

Supplementary table 6 – Homopolymer-starting site according the 5' base of the differentially represented miRNAs identified in HB4a and C5.2 by SOLiD and PGM, P= 0,4126

Beginning the base homopolymer	PGM	SOLiD
–	3	6
0	1	0
1	2	0
13	0	1
14	1	0
2	1	0
6	1	1
8	0	1

Supplementary table 7 – Size miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P=0,914

Size miRNA (number of nucluotides)	PGM	SOLiD
18	2	1
19	1	2

20	3	2
21	3	5
22	25	20
23	9	6
24	2	3

Supplementary table 8 – Size homopolymer miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P= 0,7615

Supplementary table 9 – Number of homopolymers in the miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P=0,2328

Homopolymers amount	PGM	SOLiD
0	35	25
1	9	11
2	1	3

Supplementary table 10 – Number of loops in the miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P= 0,2676

Loops amount	PGM	SOLiD
0	2	3
1	28	23
2	13	13
3	2	0

Supplementary table 11 – Base of homopolymers in the miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P=0,441

Base homopolymer	PGM	SOLiD

–	35	25
a	3	3
c	2	6
g	3	4
u	3	4

Supplementary table 12 – Beginning the ase of homopolymers in the miRNAs outliers identified in HB4a or C5.2 by SOLiD and PGM, P= 0,7636

Beginning the base homopolymer	PGM	SOLiD
–	35	25
0	2	1
1	2	2
10	0	1
11	0	1
13	0	2
14	1	1
15	0	2
2	1	2
4	1	0
5	1	2
6	2	2
8	1	1

5. References

Alvarez-Garcia I, Miska EA. MicroRNA functions in animal development and human disease. Development 2005; 132:4653-62.

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009; 36:215-33.

Carraro, D.M., Ferreira, E.N., de Campos Molina, G., Puga, R.D., et al., Poly (A)+ transcriptome assessment of ERBB2-induced alterations in breast cell lines. PLoS One 2011, 6, e21022.

Chatterjee A, Leichter AL, Fan V, Tsai P, Purcell RV, Sullivan MJ, Eccles MR. A cross comparison of technologies for the detection of microRNAs in clinical FFPE samples of hepatoblastoma patients. Scientific Reports **5**: 10438 (2015).

Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Brief Bioinform 2009; 10:490-7.

Friedländer MR, Chen W, Adamidi C, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008; 26:407-15.

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012 40

Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010; 466:835-40.

Hafner M, Renwick N, Brown M, Mihailovic´ A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. 2011. RNA-ligase dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA 17: 1697–1712.

Hager G. Footprints by deep sequencing. Nat Methods 2009; 6:254-5.

Harris, R. a, Eichholtz, T.J., Hiles, I.D., Page, M.J., O'Hare, M.J., New model of ErbB-2 over-expression in human mammary luminal epithelial cells. Int. J. Cancer 1999, 80, 477–84.

Huang J, Hao P, Chen H, Hu W, Yan Q, Liu F, Han ZG. Genome-wide identification of Schistosoma japonicum microRNAs using a deepsequencing approach. PLoS One 2009; 4:e8206.

Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. Nat Rev Mol Cell Biol 2010; 11:252-63.

Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA

biogenesis, function and decay. Nat Rev Genet 2010; 11:597-610.

Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75:843-54.

Leshkowitz D, Horn-Saban S, Parmet Y, Feldmesser E. Differences in microRNA detection levels are technology and sequence dependent. RNA. 2013,19:527-38.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. Nat Methods 6: 474–476.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012 30:434-9.

Lu YC, Smielewska M, Palakodeti D, et al. Deep sequencing identifies new and regulated microRNAs in Schmidtea mediterranea. RNA 2009; 15:1483-91.

McKernan K, Blanchard A, Kotler L, Costa G. Reagents, methods, and libraries for bead-based sequencing. US patent application 20080003571, 2006.

Meire E, Levy A, Benjamin H, et al. Discovery of microRNAs and other small RNAs in solid tumors. Nucl Acids Res 2010; 38:6234-46.

Nassirpour R, Mathur S, Gosink MM, Li Y, Shoieb AM, Wood J, O'Neil SP, Homer BL, Whiteley LO. Identification of tubular injury microRNA biomarkers in urine: comparison of next-generation sequencing and qPCR-based profiling platforms. BMC Genomics 15:485, 2014.

Novelli G, Predazzi IM, Mango R, Romeo F, Mehta JL. Role of genomics in cardiovascular medicine. World J Cardiol. 2010 2:428–436.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 13:341.

Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135–1145. doi: 10.1038/nbt1486

Stamps, A.C., Davies, S.C., Burman, J., O'Hare, M.J., Analysis of proviral integration in human mammary epithelial cell lines immortalized by retroviral infection with a temperature-sensitive SV40 T-antigen construct. Int. J. Cancer 1994, 57, 865–74.

Stark, M. S., Tyagi, S., Nancarrow, D. J., Boyle, G. M., Cook, A. L., Whiteman, D. C., Parsons, P. G., Schmidt, C., Sturm, R. A., and Hayward, N. K. (2010). Characterization of the melanoma mirnaome by deep sequencing. PLoS One, 5(3):e9685.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008, 455: 1061–1068.

Tian G, Yin X, Luo H, Xu X, Bolund L, Zhang X, Gan SQ, Li N. 2010. Sequencing bias: Comparison of different protocols of microRNA library construction. BMC Biotechnol 10: 64.

Ul Hussain M. Micro-RNAs (miRNAs): genomic organisation, biogenesis and mode of action. Cell Tissue Res. 2012 349:405-13.

Van Dijk EL, Auger H, Jaszczyszyn, Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics 2014; 30:418-426.

Van Nieuwerburgh F, Soetaert S, Podshivalova K, Ay-Lin Wang E, Schaffer L, Deforce D, Salomon DR, Head SR, Ordoukhanian P.2011. Quantitative bias in Illumina TruSeq and a novel post amplification barcoding strategy for multiplexed DNA and small RNA deep sequencing. PLoS ONE 6: e26969.

Wei B, Cai T, Zhang R, et al. Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (Triticumaestivum L.) and Brachypodiumdistachyon (L.) Beauv. Funct Integr Genomics 2009; 9:499-511.

Willenbrock H, Salomon J, Søkilde R, Barken KB, Hansen TN, Nielsen FC, MøllerS, Litman T. Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing. RNA 15:2028-2034, 2009.

Yang X, Chockalingam SP, Aluru S. A survey of error-correction methods for next-generation sequencing. Brief Bioinf 2013;14:56–66

Yu J, Wang F, Yang GH, et al. Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. Biochem Biophys Res Commun 2006; 349:59-68.

Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011 38:95-109.

miRBase. Browse miRBase by species 2012. Available from: <URL:http://www.mirbase.org/cgi-bin/browse.pl?org=hsa>[2012 dez 03].