Published in final edited form as:

Cancer Genet. 2016 March ; 209(3): 57–69. doi:10.1016/j.cancergen.2015.12.001.

Overview of recurrent chromosomal losses in retinoblastoma detected by low coverage next generation sequencing

A.J. García-Chequer^a, A. Méndez-Tenorio^b, G. Olguín-Ruiz^b, C. Sánchez-Vallejo^b, P. Isa^c, C.F. Arias^c, J. Torres^a, A. Hernández-Angeles^a, M.A. Ramírez-Ortiz^d, C. Lara^d, M.L. Cabrera-Muñoz^d, S. Sadowinski-Pine^d, J.C. Bravo-Ortiz^e, G. Ramón-García^e, J. Diegopérez-Ramírez^e, G. Ramírez-Reyes^e, R. Casarrubias-Islas^e, J. Ramírez^f, M.A. Orjuela^g, and M.V. Ponce-Castañeda^{a,*}

^aUnidad de Investigación Médica en Enfermedades Infecciosas, Centro Médico Nacional SXXI, Instituto Mexicano del Seguro Social, México D.F., Mexico

^bLab. Bioinformática Genómica, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, México D.F., Mexico

^cInstituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

^dHospital Infantil de México Federico Gómez, México D.F., Mexico

^eHospital de Pediatría, CMN SXXI, Instituto Mexicano del Seguro Social, México D.F., Mexico

^fUnidad de Microarreglos, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, México D.F., Mexico

^gColumbia University, New York, USA

Abstract

Genes are frequently lost or gained in malignant tumors and the analysis of these changes can be informative about the underlying tumor biology. Retinoblastoma is a pediatric intraocular malignancy, and since deletions in chromosome 13 have been described in this tumor, we performed genome wide sequencing with the Illumina platform to test whether recurrent losses could be detected in low coverage data from DNA pools of Rb cases. An *in silico* reference profile for each pool was created from the human genome sequence GRCh37p5; a chromosome integrity score and a graphics 40 Kb window analysis approach, allowed us to identify with high resolution previously reported non random recurrent losses in all chromosomes of these tumors. We also found a pattern of gains and losses associated to clear and dark cytogenetic bands respectively. We further analyze a pool of medulloblastoma and found a more stable genomic profile and previously reported losses in this tumor. This approach facilitates identification of recurrent deletions from many patients that may be biological relevant for tumor development.

*Corresponding author. vponce@ifc.unam.mx.

Supplementary data

Supplementary data related to this article can be found online at doi:10.1016/j.cancergen.2015.12.001.

Keywords

Pediatric tumors; retinoblastoma; medulloblastoma; next-generation sequencing; low coverage; losses

Introduction

Retinoblastoma (Rb) is an intraocular malignancy that develops in early childhood and originates in the retina. About 60–70% of all cases are unilateral and 30–40% bilateral and unfortunately surgical removal of the eye is often the first line of treatment (1). Retinoblastoma is considered a robust clinical model for genetic predisposition to development of cancer, because offspring of patients with bilateral Rb are also affected by the disease exhibiting a dominant autosomic pattern of inheritance. In addition, patients with bilateral Rb also have a significantly higher lifetime risk to develop secondary malignancies. Early cytogenetic analyses of this tumor demonstrated recurrent losses of the long arm in chromosome 13 (2–6), and these gross deletions allowed the localization and cloning of the retinoblastoma susceptibility gene RB1, the first tumor suppressor gene discovered (7,8).

Genetic alterations are hallmarks of malignant processes. In particular recurrent genetic alterations can be detected across different patients, and such recurrence is an indication that these are driving events for tumor development (9). Recurrent losses may indicate the presence of tumor suppressor genes within the lost regions, as was the case for RB1, while gained or amplified loci may translate in increased expression of amplified genes, often functioning as oncogenes like ERBB2 (also known as Her-2/neu) gene in breast cancer (10,11).

Our ability to detect oncogenic driver genes whether lost or gained has been limited by the resolution of available methods to detect these key recurrent gains or losses. In retinoblastoma recurrent chromosome gains and losses other than in chromosome 13, have been described although with variable frequencies using karyotype, FISH (Fluorescence In Situ Hybridization), CGH (Comparative Genome Hybridization), SNPs (Single Nucleotide Polymorphism), qPCR and whole genome sequence (12–19), including gains at 1q, 6p, 17q and 19q and losses at 16q and X, 2p, 5q. Thus, more oncogenes and tumor suppressor genes are thought to be or have been located in those altered regions (20–22).

High density data technologies like microarrays and massive parallel sequencing are new powerful tools able to detect many of these variations in gene copy number (23–25). Next generation sequencing (NGS) is a high throughput technology used with increasing frequency given to decreasing costs. NGS allows the sequencing of complete genomes through the making and sequencing of DNA libraries that are overlapping discrete fragments of the genome. The sequences obtained are a large collection of so called “reads”. According with the length and number of reads obtained per sample, it is nearly possible to reconstruct complete human genomes. Depth and coverage depend on the number of reads obtained in relation to the genome length, and it refers to the number of times a given region has been covered by different reads. Depending on the depth of coverage of massive sequencing data, different types of analyses requiring different resolution can be performed with NGS data,

including genome reconstruction, determination of mutations, polymorphisms, splicing events, isoforms analysis, gains and losses and translocations (26–28).

Because this technology is relatively new, methods for analysis, bioinformatic algorithms and validation to detect gains and losses are still in development (28–30), and it has not yet been thoroughly established what are the limits on the depth of sequencing data that can be useful for the identification of recurrent gains and losses (31).

Since retinoblastoma is a tumor that has been widely studied in terms of recurrent gains and losses, we chose to utilize 8 Rb cases as biological replicas to validate a procedure using low coverage-massive sequencing data to detect recurrent losses. To do this we used for sequencing pooled DNA from different retinoblastoma cases, assuming that detection of losses in a pool indicates recurrence, since all the tumors in the pool must contribute to the loss in order to be detected. In contrast gains detected cannot be called recurrent with this approach because it is impossible to discriminate the individual contribution for gains in such a pool. We also developed a bioinformatic strategy to replace the complexities and costs of sequencing a normal tissue sample with an *in silico* reference for each tumor pool sequenced. This allowed us to use log₂ ratios of the number of reads between tumor and reference to detect and quantify gains and losses. We further tested this approach analyzing the profile of pooled samples of medulloblastoma, another malignant primitive neuroectodermal tumor that is histomorphologically similar to retinoblastoma. In both tumors we found losses and gains previously reported, in retinoblastoma many more gross recurrent and non random losses were identified in most chromosomes, while in medulloblastoma we found very few gross alterations. Interestingly in retinoblastoma we found an association between gross deletions and dark cytogenetic bands and gross amplifications and the corresponding clear cytogenetic bands.

Materials and methods

DNA samples

DNA from eight retinoblastoma and five medulloblastoma tumor samples was used for the next generation sequencing analyses. Two DNA pools of four Rb samples each, as technical replicas, and one DNA pool containing five Mb patients, were sequenced. Each Rb pool was gender balanced containing 2 boys and 2 girls and the Mb pool consisted of 4 boys and 1 girl (see Table 1 for further details).

All patients were newly diagnosed and tumor samples were collected at time of surgery, prior to any adjuvant therapy, from children diagnosed and treated at Hospital de Pediatría from Instituto Mexicano del Seguro Social (IMSS) and Hospital Infantil de México Federico Gómez from Secretaría de Salud in Mexico City. Retinoblastoma samples were collected as part of a separate study which included demographic and clinical data collection as previously described (32). Tumor tissues were collected under informed written consent from their parents and as part of studies approved by the Scientific and Ethics Review Boards from each participating institution. Patient characteristics are shown in Table 1. DNA was extracted from the fresh frozen tumor tissues by a non-organic method using Qiagen

and 5PRIME (DNeasy Blood and Tissue Kit, ArchivePure DNA Blood kit) following the manufacturer directions.

Illumina sequencing and expression microarrays

Sequencing was performed at the sequencing core facility of the National University of Mexico (UNAM) located at the Biotechnology Institute in Cuernavaca, Mexico using the Illumina Genome Analyzer IIx (Illumina). For sequencing 5 µg of DNA, 1.2 µg DNA from each retinoblastoma per pool, and 1 µg DNA from each medulloblastoma per pool were used to make approximately two hundred base-pair sized libraries using Illumina's Genomic DNA sample prep kit. Libraries were loaded in separate lanes of a flow cell, and sequencing was performed through 35 cycles of single base pair extensions. Image analysis and capture was done using goat.py, and base calling was carried out using Bustard.pl, both programs included in the Off-Line Basecaller Software v1.8.0 (Illumina).

Expression microarray data were generated at the microarray facility of the National University of Mexico (UNAM) located at the Instituto de Fisiología Celular in Mexico City. Total RNA was obtained using Trizol reagent by Ambion according to manufacturer directions, from short term primary cultures established from enucleated eyes and propagated for seven days in RPMI 15% FBS. In house double channel platform was used and microarrays were printed with a human 50-mer oligo library set "A" from MWG BiotechOligo Sets, as a reference RNA from *S. cerevisiae* harvested at growth logarithmic phase was used. The microarray contains approximately 10,000 gene specific oligonucleotide probes representing the best-annotated genes from human. Acquisition and quantification of array images was performed with aScanArray 4000 scanner with its accompanying software ScanArray 4000 from Packard BioChips. All images were captured using 65% PMT gain, 70–75% laser power and 10 µm resolution at 50% scan rate. For each spot the Cy3 and Cy5 density mean value and the Cy3 and Cy5 background mean value were calculated with software ArrayPro Analyzer from Media Cybernetics. After normalization using Local Lowes method, log ratio expression from tumor relative to reference was calculated for each gene. Gene expression was determined as significant over or under expressed using the Significance Analysis of Microarray (SAM) algorithm from the MeV suite of programs setting the threshold for significance with a log fold change value larger than 2 or lower than -2 for selecting the over and under expresses genes respectively and a false discovery rate of 0.05. Data sets are uploaded at GEO-NCBI database in: <http://www.ncbi.nlm.nih.gov/geo/> accession number GSE11488.

Filtering of the data

All reads had base call accuracy of $Q > 30$. The original sequencing data were filtered in two steps, first low-quality reads and low complexity sequences were removed, then frequencies of repeated reads were calculated allowing to collapse in one, all repeated reads. For Rb pool 1 the raw and filtered number of reads were respectively 31,019,035 and 22,467,328; for Rb pool 2, 33,895,793 and 29,218,515 and for Mb pool 31,976,350 and 28,555,471.

***In silico* references**

In order to analyze sequencing data from the samples to identify gains and losses, a reference profile of sequencing data for each sample was calculated using an *in silico* technique. The whole human genome sequence GRCh37p5 was downloaded from the Genome Reference Consortium. Then a collection of short sequences of the same size as those generated by the Illumina Genome Analyzer (35 mer) were randomly sampled from this genomic sequence and considered *in silico* reads. As the real samples consisted of pools of genomic material from both female and male individuals, their proportion in each pool was considered for a representative sampling of the X and Y chromosomes in the *in silico* sets. Each set was then filtered using the same method which was used for the original sequencing data as mentioned before. The number of *in silico* reads after filtering was 22,102,888 for Rb pool 1, 29,215,286 for Rb pool 2 and 28,459,131 for Mb pool.

Reads treatment

The treatment applied to both the *in silico* and the real reads was the same. After filtering, both *in silico* and real sets were then mapped to the GRCh37.p5 genome sequence with Bowtie2 (33). A Perl program was written to keep the reads that map only once to the Human Genome and their positions were stored for each data set. For the comparison between the *in silico* reads and *in vitro* reads already mapped, $\log_2$ ratios values were used.

The log ratio value is defined as the log base 2 of the relation between the number of reads in a tumor sample and the number of reads in a reference sample.

$$\log_2 = \frac{\text{Number of reads in Sample X}}{\text{Number of reads in Reference X}}$$

The $\log_2$ ratio gives a symmetric relationship between two sets of values facilitating the interpretation of the data. For example, a 2 value will be equivalent to a -2, and this will correspond to a relation of 4:1 and 1:4 respectively or a fold change of 4.

ChromDraw

ChromDraw.exe is a computational program that draws each chromosome and all its cytogenetic bands. This graphic visualization software was written by one of the authors of the present work (Méndez-Tenorio, PhD) with the purpose of facilitating identification of losses and gains in each chromosome by analyzing the comparison between one sample and its reference using the $\log_2$ ratio value. This software is available upon request.

Integrity score

The integrity score was calculated as a relation between the number of reads in a chromosome and the length of that chromosome, in units of reads/kilobase. The integrity score was used to obtain an overview of gross losses and gains in all the chromosomes. This score is calculated as follows:

$$score_{Chri} = \frac{\text{number of reads in chromosome } i}{\text{length of chromosome } i} \left[= \right] \frac{\text{reads}}{\text{kb}}$$

Using this integrity score we are able to determine in a simple assay if there is any chromosome in the sample with less or more reads compared to its reference.

Results

Overview DNA data sets

The original and raw data are uploaded at the European Nucleotide Archive and can be found in <http://www.ebi.ac.uk/ena/data/view/PRJEB6630>. Genome coverage was defined as the number of times a given region has been sequenced by different reads. The depth of coverage was calculate by the formula $c = (m * n)/q$ where m is the number of reads per sample, n the average length of the reads and q the length of the genome (34). By assuming an average human genome length of 2,988,355,349 base pairs (bp), then the depth of coverage obtained from filtered reads was very low $<1 \times$ in the three pools, for Rb pool 1 = 0.36, for Rb pool 2 = 0.39 and for the Mb pool = 0.37.

In silico genome references

In order to assess losses and gains each read is considered as a probe and the number of reads per region of each chromosome must then be compared with a reference, which usually is a genome from non cancer or healthy tissues from which the tumor originates (35). To circumvent the complexities, challenges and cost entailed in having a non-tumor reference genome matching each individual tumor, we created an *in silico* reference genome. This approach considers the usefulness of a reference that is not derived from “normal” tissue from which a tumor arises in the same individual, but rather reflects the magnitude of the data obtained in the sequenced sample. This is based on the assumption that massive sequencing data are derived from a random process and the human genome from which the reference is taken originates from an apparently healthy individual.

To create each reference genome, first the whole human genome sequence GRCh37p5 was downloaded from the Genome Reference Consortium and the sequence of each chromosome was linked to a single linear thread considering a 3:1 ratio for the X:Y chromosomes for Rb pools and 3:2 for Mb pool, according to the gender composition of each pool mentioned before. This linear genome sequence was then randomly sampled at any given position, taking from it on but not cutting, the following short sequences of 35 nt, that is copying small sequences the same size as those generated by the Illumina Genome Analyzer (35 mers). The process allowed repetition until reaching the matching number of reads obtained from the real pool. With an *in silico* reference genome for each sequencing set of pooled DNA we were able to use reads ratios between each tumor pool and its reference to calibrate and measure gains and losses (see section “Chromosome losses and gains ...” for details).

Chromosome integrity score

We first developed an integrity score to have a panoramic view of losses and gains in each chromosome across the genome using NGS data. We defined the integrity score as the relationship between the number of hits (reads hitting a unique place in a chromosome) and the length of that chromosome, with the integrity score's units defined as hits/kb. We reasoned that if the number of reads yielded by the NGS is proportional to the chromosome length, then the integrity score of reference and tumor will detect gains and losses. To test this we first analyzed the relationship between the chromosome length and the corresponding number of uniquely mapped reads. The integrity score was calculated for all the chromosomes in the six data samples, three for the reference samples and three for the pooled tumor samples. By plotting the size of each chromosome in millions of bp against the number of reads obtained for each chromosome; we were able to confirm linearity. Figure 1 shows the plot for a reference sample as an example.

We obtained an r^2 value of 0.93 in each of the three reference samples, while in both Rb tumor samples r^2 was 0.83. This difference is not surprising since the tumor samples are samples with expected alterations in DNA content, and are comprised of pooling patients with varying characteristics (see Table 1 for clinical data) and thus biological variability is expected. For the Mb pool the correlation factor was 0.93 the same as the references suggesting less variability in this tumor pool. As expected ChrY is the most dispersed element for these correlations (Figure 1), and when this point is removed the square correlation factor (r^2) increases from 0.93 to 0.94 in the *in silico* references, from 0.83 to 0.84 in the Rb tumor samples and from 0.87 to 0.95 in Mb. In Figure 1B the Integrity Score is plotted as bars and a marked difference for chromosome Y is also observed, this condition is present in all tumors and reference samples. ChrY represents an exception among the chromosomes due to its smaller size, elevated number of repetitive sequences, and relative scarcity of genes compared with others chromosomes. ChrY was therefore excluded from subsequent analyses related to the integrity score.

A general, although low power view of each chromosome's integrity score across the three pools of tumor DNA is presented in Figure 2. The integrity score was adjusted to the reference (subtracting sample minus *in silico* integrity scores), standard deviation was calculated and limits were established at 2*standard deviations above and below zero. All the chromosomes that exceed the limits (2*sd) are those that have the biggest losses or gains, what this score does not tell is where within the chromosomes these losses occur. The retinoblastoma pool 1 is the sample that seems to have losses in more Chrs, since the integrity score exceeds the limits in 8 chromosomes as losses (Chr2, Chr3, Chr4, Chr5, Chr6, Chr8, Chr13, Chr18 and ChrX) and in one as gains (Chr19). The retinoblastoma pool 2 shows important losses just in two chromosomes (Chr4 and Chr13) and gains in one (Chr22). Finally medulloblastoma pool shows losses also in two chromosomes (Chr6 and ChrX) however losses observed in ChrX in Mb are more pronounced than those found in Rb pools. Using the integrity score in a straightforward analysis, we were able to determine whether any tumor chromosomes had fewer or more probes compared with the reference.

Chromosome losses and gains analysis by “windows”

In order to easily identify patterns of gains and losses within the chromosomes with better resolution and accuracy, we designed a system to scan ratios locally and implemented the use of graphics that allow to relate plotted ratios between *in silico* reference and pooled samples data with cytogenetic bands (ideogram) at each chromosome (ChromDraw). Each chromosome sequence was divided into fragments of the same length called “windows” and slided throughout the genome. Mapped reads in each window for both the *in silico* reference and the tumor samples were measured. Two different window sizes were initially used, 10 kb and 40 kb; however since we found greater dispersion in the data using the 10 kb windows, we decided to use the 40 kb window. A total of 78, 510 windows were obtained for each sample, with an average of 261 reads/window, a maximum of 2656 reads and a minimum of 0 reads per window. Locally scanned ratios between the *in silico* reference and pooled samples data were plotted above the corresponding chromosomal ideogram depicting cytogenetic bands, adjusting the size of each chromosome ideogram to the size (length) of the plotted ratios. Establishing upper and lower limits this graphic display allows visual analysis and association of losses or gains detected with the NGS data to the corresponding cytogenetic bands of each chromosome, facilitating a genome wide view of losses and gains. See Supplementary data at (36) for the individual view of each chromosome and corresponding genes distribution in the Maps File. With this approach we first focused the analysis on Chr13 to define the extent of losses we successfully detected with the integrity score mentioned before and results are presented in Figure 3. Consistent with previous reports, important losses were observed at the q arm of the Chr13 of both pools which includes the retinoblastoma locus (13q14.2). This loss spans well beyond the retinoblastoma locus including more than 11 cytogenetic bands from 13q14.2 to 13q31.3 in pool 1. Other losses from 13q33.1 to 13q33.3 previously reported associated to Rb were also identified (36). Thus, “windows” analysis confirmed results obtained with the integrity score, and revealed the cytogenetic location and extension of recurrent losses at the RB locus with higher accuracy. Chr19 was the only chromosome in Rb pool 1 and Chr22 in pool 2 showing wide spread gains along each chromosome in agreement to their corresponding integrity score. Other gains are in Chr1, 6, 7, 9, 11, 12, 16, 17, 19, and 22 although they were of lower magnitude and more discontinuous when compared with the losses observed in Chr13 or ChrX. Since we know that losses detected with this pooling approach reveal recurrent losses but not recurrent gains among the pooled tumors, we then focused the analysis on recurrent losses in all other chromosomes finding indeed, recurrent losses previously reported by other authors in all chromosomes (14,37,38). Patterns of similar losses between Rb pools 1 and 2 can be appreciated across the chromosomes, suggesting that recurrent losses are not random, and even though the patterns between Rb pools can be discerned, they tend to be larger in pool 1, and although scattered more gained regions are also present in pool 1. An important difference between pools 1 and 2 could account for this, namely differences in disease spread as 2 out of 4 cases in pool 1 had more advanced clinical stage as shown in Table 1, including optic nerve involvement and extraocular spread, while all tumors in pool 2 did not have either. From a more detailed analysis an interesting and unreported pattern emerged, consisting of an association although loose, between losses and gains in the plotted ratios with the corresponding dark and clear cytogenetic bands of the chromosomal ideograms. Table 2 shows the chromosomes, arm and cytogenetic bands on which we found more consistency in

this association. Figure 4 also illustrates Chr 7 from Rb pool 1 as the chromosomal element that display with more clarity the pattern gain-clear cytogenetic bands, and Chr11 with the pattern loss-dark cytogenetic bands (see also Maps File in (36)).

The windows profile of all chromosomes across the three pools is presented in Figure 5, which shows a more uniform and compacted profile for the Mb pool, suggesting again that genomes are more stable in Mb than in Rb. The observed displacement of the data density in favor of ChrY in Mb confirms the gender composition of the patients in this pool, as noted before, this pool is composed by four boys and one girl (see Table 1). Also of notice is the scarcity of gains and losses identified in Mb pool compared with those alterations reported in the literature for this tumor. The displacement of the density data in favor of ChrX in the reference data confirms important losses at ChrX in both Rb and Mb pools. Furthermore the losses at Chr13 identified by the integrity score and by the window analysis in the Rb pools, were also detectable in these all-chromosomes profiles as a displacement of the density data at Chr13 in the Rb but not in the Mb profiles. These observations validate the utility of the presented simplified approach to study gains and recurrent losses in multiple cases in the same sequencing run.

The medulloblastoma and retinoblastoma tumors shared losses

The integrity score analyses showed shared gross losses between Mb and Rb at ChrX, and a more genome wide homogeneous profile in Mb pool. Analysis of losses between Rb and Mb with plotted data using the windows analyses, revealed a recurrent loss at Chr6 spanning three cytogenetic bands including 6p22.1, 6p21.33 and 6p21.32 shown in Figure 6. It spans a region of approximately 5 Megabases from position 29,000,000 to 33,000,000. The major histocompatibility complex (MHC) is situated from position 28,510,120 to 33,480,577 and is an important region that include a group of genes encoding cell surface molecules involved in cell mediated immunity and it is extremely polymorphic. For this reason this region has been very difficult to assemble using NGS (39,40), and the shared losses observed are more likely an artifact of the mapping process employed here than real recurrent losses. Another lost region shared between Mb and Rb is situated in Chr 19 at the cytogenetic band 19q13.42. It goes from position 54,520,000 to 55,560,001 spanning 1 Mb of this chromosome and 66 genes are situated in this region (see Table 1 in (36)). This region has been previously reported to harbor a human leukocyte receptor cluster (41) and 10 of these 66 genes belong to the KIR gene family (Killer cell immunoglobulin-like receptors) these are also polymorphic and highly homologous genes, thus it is likely that these apparent losses are also an artifact of the mapping process mentioned before. Other shared losses exist in Chr17q21.3, Chr19p13.11, Chr5q13.2, and although are of smaller magnitude it will be interesting a deeper analysis on them (See Table 1 in (36)).

Validation of the next generation sequencing with expression microarray data

To determine if expression microarray data could inform us about the functional relation between the gains and losses detected, we searched for correlations between NGS data (DNA) and data previously obtained with expression microarrays (RNA). Expression microarray data were generated with a custom double-channel low density microarray platform with six Rb cases from which RNA was available: four cases from pool 1 and two

cases from pool 2 (cases F-1 and F54 Table 1). The approximately 10,000 human genes contained in the microarray were mapped against the human genome for the evaluation of the expressed genes along the chromosomes. The expression intensities from all the genes analyzed are distributed along the genome and shown in Figure 7A. A Significance Analysis of Microarrays (SAM) to detect highly expressed genes was done and only the genes significantly over expressed per chromosome are shown in Figure 7B. Chr19 and Chr17 were the chromosomes that contained the highest number of genes with significantly increased expression, and conversely Chr13 is one of the chromosomes without significantly over expressed genes, consistent with the integrity score results in Figure 7D. A list of these genes can be seen as Table 2 in (36). Finally, the data from DNA and RNA techniques were compared to analyze gains in Chr19. The results showed some agreement between DNA gains and those genes more highly expressed (Figure 7C).

Discussion

NGS as a tool for evaluating recurrent losses by using pooled samples

Our results indicate that low coverage NGS data can be very useful to detect recurrent losses pooling samples from different patients in one sequencing run, and surprisingly with very high resolution. Even though gains cannot be defined as recurrent with this approach, nonetheless complementing low coverage sequencing data with microarray data allows the identification of gains with functional consequences. We focused the analysis on recurrent losses because these events more likely are cancer driver events in all retinoblastomas studied here. In contrast, over expressed genes located in regions with gains in the context of this work, suggest that detected gains may have functional consequences but cannot be called recurrent events in this group of retinoblastomas. Patterns of losses and gains in RB pools and its association with cytogenetic banding not reported before, provoke interesting questions regarding the topology and mechanisms of gains and losses in a cellular context without retinoblastoma protein (42). The presented strategy gives a broad topographic and genome wide view with a density component of gains and recurrent losses, allowing further to prioritize the regions and genes to be studied and tested. Broad patterns of losses and gains between Rb and Mb are clearly different although there are many smaller shared regions that call for a more detailed analysis.

Since recurrent alterations in retinoblastoma do not occur in 100% of the cases, as reported by other researchers, we think that specific patterns of recurrent events or profiles can be attributed to inter tumor heterogeneity. It is also conceivable that specific patterns of recurrent alterations may correspond to relevant biological or clinical features, such as differences in response to treatment, invasion or metastatic potential, and it is possible that these recurrent alterations may differ between populations. Thus a potentially useful application of this approach would be the pooling of patients with clinical relevant phenotypes to search for recurrent losses and thus pointing to loci and genes involved in the phenotype of interest.

Conclusion

We found robust evidence of recurrent losses in retinoblastoma consistent with what is reported in the literature, an association of gains and losses with clear and dark cytogenetic bands, and genome wide non random patterns of recurrent losses in the two pools of Rb. The significance and utility of this work relies on the use of (1) low coverage NGS data, (2) pooled malignant tumor DNA samples, (3) an *in silico* reference for each data set to effectively map gains and recurrent losses and (4) the development of visual graphics allowing a resolution of 40 Kb for detecting gains and losses. NGS of pooled samples is a less expensive alternative to search for recurrent losses than current methods for example karyotyping, that requires tumor by tumor analysis to search for recurrent events since many patients can be studied in a single NGS experiment; this is particularly relevant in settings with constrained economic resources, or tumor samples not suitable for karyotyping. Our results indicate that low coverage NGS data can be useful to detect recurrent losses with high resolution, and that combined with expression microarray data, NGS allows the identification of gains with functional consequences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by CONACYT SALUD-2003-C01-83, IMSS 2002/111, 2007-571, and NIH CA167833, CA98180. JT is a recipient of an Exclusivity Scholarship from Fundación IMSS, Mexico. We thank L. Chávez-González, S. Guzmán-León and J.L. Santillán-Torres for technical assistance in the microarray determinations.

References

1. Abramson DH. Retinoblastoma: diagnosis and management. *CA Cancer J Clin.* 1982; 32:130–140. [PubMed: 6804033]
2. Lele KP, Penrose LS, Stallard HB. Chromosome deletion in a case of retinoblastoma. *Ann Hum Genet.* 1963; 27:171–174. [PubMed: 14081487]
3. Wilson MG, Towner JW, Fujimoto A. Retinoblastoma and D-chromosome deletions. *Am J Hum Genet.* 1973; 25:57–61. [PubMed: 4119334]
4. Francke U, Kung F. Sporadic bilateral retinoblastoma and 13q-chromosomal deletion. *Med Pediatr Oncol.* 1976; 2:379–385. [PubMed: 1004381]
5. Pogosianz H, Kuznetsova LE. Nonrandom chromosomal changes in retinoblastomas. *Arch Geschwulstforsch.* 1986; 56:135–143. [PubMed: 3707294]
6. Orye E, Delbeke MJ, Vandenabeele B. Retinoblastoma and long arm deletion of chromosome 13. Attempts to define the deleted segment. *Clin Genet.* 1974; 5:457–464. [PubMed: 4854145]
7. Friend SH, Bernards R, Rogelj S, et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature.* 1986; 323:643–646. [PubMed: 2877398]
8. Dryja TP, Rapaport JM, Joyce JM, et al. Molecular detection of deletions involving band q14 of chromosome 13 in retinoblastomas. *Proc Natl Acad Sci U S A.* 1986; 83:7391–7394. [PubMed: 2876425]
9. Dutt A, Beroukhis R. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol.* 2007; 19:43–49. [PubMed: 17133111]
10. Slamon DJ, Clark GM, Wong SG, et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science.* 1987; 235:177–182. [PubMed: 3798106]

11. Ross JS, Fletcher JA. The HER-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy. *Stem Cells*. 1998; 16:413–428. [PubMed: 9831867]
12. Herzog S, Lohmann DR, Buiting K, et al. Marked differences in unilateral isolated retinoblastomas from young and older children studied by comparative genomic hybridization. *Hum Genet*. 2001; 108:98–104. [PubMed: 11281459]
13. Corson TW, Gallie BL. One hit, two hits, three hits, more? Genomic changes in the development of retinoblastoma. *Genes Chromosomes Cancer*. 2007; 46:617–634. [PubMed: 17437278]
14. van der Wal JE, Hermesen MA, Gille HJ, et al. Comparative genomic hybridisation divides retinoblastomas into a high and a low level chromosomal instability group. *J Clin Pathol*. 2003; 56:26–30. [PubMed: 12499428]
15. Lillington DM, Kingston JE, Coen PG, et al. Comparative genomic hybridization of 49 primary retinoblastoma tumors identifies chromosomal regions associated with histopathology, progression, and patient outcome. *Genes Chromosomes Cancer*. 2003; 36:121–128. [PubMed: 12508240]
16. Gilbert F, Potluri VR, Short MP, et al. Retinoblastoma, chromosome abnormalities and oncogene expression. *Ophthalmic Paediatr Genet*. 1987; 8:3–10. [PubMed: 3295640]
17. Ganguly A, Nichols KE, Grant G, et al. Molecular karyotype of sporadic unilateral retinoblastoma tumors. *Retina*. 2009; 29:1002–1012. [PubMed: 19491728]
18. Gratiás S, Schuler A, Hitpass LK, et al. Genomic gains on chromosome 1q in retinoblastoma: consequences on gene expression and association with clinical manifestation. *Int J Cancer*. 2005; 116:555–563. [PubMed: 15825178]
19. Zhang J, Benavente CA, McEvoy J, et al. A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature*. 2012; 481:329–334. [PubMed: 22237022]
20. Theriault BL, Dimaras H, Gallie BL, et al. The genomic landscape of retinoblastoma: a review. *Clin Experiment Ophthalmol*. 2014; 42:33–52. [PubMed: 24433356]
21. Orlic M, Spencer CE, Wang L, et al. Expression analysis of 6p22 genomic gain in retinoblastoma. *Genes Chromosomes Cancer*. 2006; 45:72–82. [PubMed: 16180235]
22. Conkrite K, Sundby M, Mu D, et al. Cooperation between Rb and Arf in suppressing mouse retinoblastoma. *J Clin Invest*. 2012; 122:1726–1733. [PubMed: 22484813]
23. Hayes JL, Tzika A, Thygesen H, et al. Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation. *Genomics*. 2013; 102:174–181. [PubMed: 23598253]
24. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*. 2011; 470:59–65. [PubMed: 21293372]
25. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*. 2010; 11:685–696. [PubMed: 20847746]
26. Koboldt DC, Steinberg KM, Larson DE, et al. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013; 155:27–38. [PubMed: 24074859]
27. Morozova O, Marra MA. From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. *Biochem Cell Biol*. 2008; 86:81–91. [PubMed: 18443621]
28. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*. 2008; 40:722–729. [PubMed: 18438408]
29. Teo SM, Pawitan Y, Ku CS, et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*. 2012; 28:2711–2718. [PubMed: 22942022]
30. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*. 2009; 10:80. [PubMed: 19267900]
31. Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*. 2009; 6:99–103. [PubMed: 19043412]
32. Ramírez-Ortiz MP-CM, Cabrera-Munoz ML, Medina-Sansón A, et al. Diagnostic delay and socio-demographic predictors of stage at diagnosis and mortality in unilateral and bilateral retinoblastoma. *Cancer Epidemiol Biomarkers Prev*. 2014; 23:784–792. [PubMed: 24521997]

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9:357–359. [PubMed: 22388286]
34. Brown, SM., editor. *Next-Generation DNA Sequencing Informatics*. New York, USA: Cold Spring Harbor Laboratory Press; 2015.
35. Pinkel D, Seagraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*. 1998; 20:207–211. [PubMed: 9771718]
36. Garcia-Chequer AJ, Jaimes-Diaz H, Ponce-Castaneda MV. Illumina next generation sequencing data and expression microarrays data from retinoblastoma and medulloblastoma tissues. *Data Brief*. 2015 (in press).
37. Mol BM, Massink MP, van der Hout AH, et al. High resolution SNP array profiling identifies variability in retinoblastoma genome stability. *Genes Chromosomes Cancer*. 2014; 53:1–14. [PubMed: 24249257]
38. McEvoy J, Nagahawatte P, Finkelstein D, et al. RB1 gene inactivation by chromothripsis in human retinoblastoma. *Oncotarget*. 2014; 5:438–450. [PubMed: 24509483]
39. Gabriel C, Furst D, Fae I, et al. HLA typing by next-generation sequencing—getting closer to reality. *Tissue Antigens*. 2014; 83:65–75. [PubMed: 24447174]
40. Major E, Rigo K, Hague T, et al. HLA typing from 1000 genomes whole genome and whole exome Illumina data. *PLoS ONE*. 2013; 8:e78410. [PubMed: 24223151]
41. Wende H, Volz A, Ziegler A. Extensive gene duplications and a large inversion characterize the human leukocyte receptor cluster. *Immunogenetics*. 2000; 51:703–713. [PubMed: 10941842]
42. Cook R, Zoumpoulidou G, Luczynski MT, et al. Direct involvement of retinoblastoma family proteins in DNA repair by non-homologous end-joining. *Cell Rep*. 2015; 10:2006–2018. [PubMed: 25818292]

Figure 1.

Integrity Score for a reference sample. (A) Each point corresponds to each one of the human chromosomes, the relation between the chromosome's length (in the axis) and the number of reads from the reference sample was plotted confirming the linearity of this relationship. Chromosome Y is an outlier indicated by the arrow. (B) Integrity Score plotted as bars also shows a marked difference for chromosome Y.

Figure 2. Retinoblastoma and medulloblastoma chromosome gains and losses by their integrity score. The integrity score is shown for each chromosome in the three tumor samples. The horizontal blue lines mark the limits established for losses or gains as 2* standard deviations of the reference. (A) Retinoblastoma Pool 1; (B) Retinoblastoma pool 2 and (C) Medulloblastoma pool. All three samples are adjusted with their own reference (Pool-Ref). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figure 3.

Mostly losses found in chromosome 13 in Rb pools. Each dot in the plotted $\log_2$ ratios corresponds to a 40 kb window, values between 1 and -1 are plotted in blue; dots below -1 are shown in red meaning losses and dots above 1 in green meaning gains. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figure 4.

Correspondence between gains and light cytogenetic bands and recurrent losses and dark cytogenetic bands. Blue and pink shades across the plotted ratios and ideogram from Chr 7 and 11 show the correspondence between gains and losses with light and dark cytogenetic bands. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Figure 5.

Profiles of the 40 kb windows across the chromosomes. The $\log_2$ ratios values of all windows were plotted using a spreadsheet yielding a compact genomic profile. Above 0 plotted data from tumors, below 0 data from the reference. Arrows in Rb profiles show displacement of data density toward the reference at Chr13 indicating losses in tumor. Arrow in Mb points to ChrX showing displacement of data density toward the reference for ChrX confirming losses in the tumor pool.

Figure 6. Losses in Chr 6 shared by Rb and Mb. This loss spans 3 cytogenetic bands where the major histocompatibility complex loci are found.

Figure 7.

Relationship between DNA and RNA data in Rb. (A) Gene expression intensity from approximately 10,000 genes mapped against their respective chromosomes. (B) Number of significantly expressed genes in retinoblastoma per chromosome, Chr17 and Chr19 has the highest number of highly expressed genes. (C) A comparison between DNA gains, in upper panel blue bars identified by NGS and highly expressed genes in Chr19 by expression microarrays in lower panel red bars. (D) Chromosome Integrity Score for Rb pool 1 showing a good agreement with highly expressed genes in B indicated by red ellipses and circles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1

Clinical and demographic features of participating retinoblastoma and medulloblastoma patients included in the DNA pools

Pool	ID	Gender	Laterality	St-Jude stage
Rb 1	F86	Female	Unilateral	III-c
Rb 1	F78	Male	Unilateral	II-c
Rb 1	F37	Female	Bilateral	II-d
Rb 1	F73	Male	Bilateral	II-c
Rb 2	F54	Male	Unilateral	II-c
Rb 2	F-1	Female	Bilateral	II-c
Rb 2	F81	Female	Unilateral	II-a
Rb 2	F168	Male	Bilateral	II-a

Pool	ID	Gender	Histology
Mb	FM 1	Male	Classic IV (WHC)
Mb	FM 15	Female	Classic IV (WHC)
Mb	FM 3	Male	Classic IV (WHC)
Mb	FM 17	Male	Desmoplastic
Mb	FM 22	Male	Classic IV (WHC)

Table 2

Recurrent losses in retinoblastoma pools identified by low coverage NGS associated to cytogenetic bands

Chr	Arm	Pool	Cytogenetic bands
1	q	2	q22.1, q24.1, q24.3, q32.1, q32.3
2			
3			
4	p	2	p13, p15.1
4	q	2	q26, q28.1, q28.3
5	p	2	p15.2
5	q	2	q34
6	q	2	q12, q16.1, q16.3, q22.3, q24.1
7	p	1	p12.1, p12.3
7	q	2	q21.11, q21.13, q21.3
8	p	1	p12, p22, p23.2
8	q	2	q12.11, q21.13, q21.3, q23.3
9	p	2	p21.1, p21.3, p23
9	q	1	q31.1, q33.1
10	p	1	p14
10	q	2	q21.1, q21.3, q25.1
11	p	2	p12, p14.3
11	q	2	q22.1, q22.3
12	p	2	p11.22, p12.1, p12.3
12	q	2	q14.1, q21.1, q21.31, q21.33
14	q	2	q12, q21.1, q21.3, q31.3
15	q	1	q14, q21.1, q25.3
16	q	1	q12.2, q21, q23.1
17	p	1	p12
17	q	1	q12, q24.3
18	p	1	p11.31
18	q	1	q12.1, q12.3, q21.2, q21.32, q22.1, q22.3
X	p	2	p21.1, p21.3, p22.12
X	q	2	q23, q25