Synopsis
1. Name of the Student: Ms. Poulami Das

2. Name of the Constituent Institution: Tata Memorial Centre, Advanced Centre for Treatment Research and Education in Cancer


4. Title of the Thesis: Identification of Specific Genetic Alterations in Cervical Cancer by Genome wide LOH and Copy Number Analysis

5. Board of Studies: Life Sciences
INTRODUCTION

With about 134000 new cases and 72800 deaths annually, cancer of the uterine cervix accounts for the most frequent cancer among women in India. Globally, it is the third most common cancer and significantly contributes to cancer burden (WHO/ICO). Based on histology cervical cancer can be broadly divided into 3 subtypes- squamous cell carcinoma or cancer of the flat epithelial cells with the highest incidence (80-90%) rate followed by adenocarcinoma (10-20%) arising from glandular epithelium and mixed carcinoma (1-2%) with features of both types. HPV infection has been shown to play a critical, though not sufficient role in the etiology of cervical cancer. The relationship between exposure to HPV and the disease might vary with the infecting HPV type, viral integration status and viral load; these considerations might have profound implications on patient prognosis (Josefsson et al., 2000; Woodman et al., 2007).

Identification of individual HPV types in a population is important not only to investigate the epidemiology and clinical behavior of particular type, but as a primary screen for cervical cancer detection. Till date several HPV types have been reported; 24 of the most common types have been divided into three groups based on the severity: 15 high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82), 3 putative high-risk types (26, 53, and 66), and 6 low-risk types primarily found in genital warts and low-grade cervical lesions (6, 11, 42, 43, 44 and 70) (Munoz et al., 2003; Schmitt et al., 2006). Most of the high risk HPV (HR-HPV) infections regress spontaneously and only in about 10% cases the infection persists and progresses to high-grade cervical intraepithelial neoplasia. This generally occurs through integration of the HPV genome into the host chromosome with associated loss or disruption of E2 (Wentzensen et al., 2004). In the absence of E2 the transcriptional control on E6 and E7 is lost, which leads to immortalization and transformation of the cells (Romanczuk and Howley, 1992).
Synopsis

has been reported that transcripts derived from the integrants are more stable than those from episomal viral DNA, hence the event has been associated with a selective growth advantage for affected cells (Jeon et al., 1995; Jeon and Lambert, 1995). The integration event in most cases results in the generation of mRNAs that comprises of the viral oncogenes E6 and E7 and cotranscribed cellular sequences (Klaes et al., 1999). Another important factor associated with HPV infection is the viral load. The relationship between viral load and the disease is quiet complex; while some studies report an increase in viral load with increasing disease severity, others found either no association or a higher viral load in women with low-grade squamous intraepithelial lesion than in those with high-grade squamous intraepithelial lesion (Lilo et al., 2005; Swan et al., 1999; Woodman et al., 2007). Moreover, while considering the relationship between viral load and disease prognosis, the physical status of the virus also becomes important. All these parameters taken together could therefore provide valuable insight into the natural history of HPV infections and their relationship to disease.

Apart from infection by HPV, genetic instability is also responsible for the pathogenesis of cervical cancer. The genetic instabilities may range from point mutations, copy number changes, chromosomal rearrangements to widespread aneuploidy. Somatic alterations (present only in tumour) and not in germline (present in blood) would shed light on tumorigenesis as well as response to therapy. Next generation sequencing techniques which include whole genome sequencing, exome sequencing and transcriptome sequencing, have made such studies much more feasible. Though exons constitute only about 1% of the genome (37.6 Mb), sequencing them can yield significant information as they have been reported to harbor most variations. Also most frequent type of disease mutations are those that cause amino acid substitutions resulting from variations in exons. Further, approaches involving targeted sequencing
provide increased sequence coverage of a particular region of interest at high throughput and lower cost.

The present study involves investigation of various factors involved in cervical carcinogenesis. This included genotyping of 24 HPV types along with identification of site of viral integration in a cohort of Indian women with locally advanced cervical cancers (FIGO Stage IIB and IIIB), and determining viral copy number of two high risk HPV types - HPV16 and 18. Further, whole exome sequencing has been carried out to identify the genetic alterations observed in cervical cancer biopsies compared to matched blood in some cases. The data has been validated using different techniques.

**AIMS AND OBJECTIVES**

1. Determination of HPV status in cervical cancer biopsies.
2. Identification of integration site of the virus by Amplification of Papillomavirus Oncogene Transcripts (APOT) assay.
3. Identification of genomic alterations in cervical cancer biopsies by exome sequencing
4. Correlation of the data from the study with clinical data.

**MATERIALS and METHODS**

*Clinical Sample accrual:* Pretreatment cervical tumour biopsies, predominantly from FIGO stage IIIB as well as blood samples were obtained from patients undergoing radiotherapy alone or concomitant chemo-radiation at the Radiation Oncology Department, Tata Memorial Hospital, Mumbai, after obtaining IRB approval. A generic consent for basic research was obtained prior to obtaining the biopsies. However for the current study a consent waiver was obtained from the Hospital Ethics Committee since the samples were collected more than 10 years ago. The samples were collected in liquid
nitrogen and stored at -80°C freezer until further use. All the samples were assigned a laboratory code to maintain confidentiality.

**Processing of tumour samples:** Cryosectioning of cervical biopsies was done to determine tumour percentage as well for isolation of DNA and RNA. First, two 5µm sections of tissues were mounted on lysine coated slides for H&E staining and determining tumour percentage. Next, seven 30µm sections were collected in RLT buffer for isolation of RNA and five 30µm sections collected in STE buffer for isolation of DNA.

**Genomic DNA extraction:** DNA was extracted using standard phenol/chloroform method and quantitated using Nanodrop. For the exome sequencing study isolation of genomic DNA was done by DNA Mini kit (Qiagen) according to the manufacturer’s protocol. The integrity of DNA was checked on 0.8% Agarose gel.

**Extraction of RNA:** RNA was extracted using RNeasy Mini Kit. DNAse treatment of the RNA was done to remove any DNA contamination. RNA was quantitated using Nanodrop and its integrity was checked on 1.2% Formaldehyde gel.

**Genotyping of HPV by high throughput Luminex assay:** Genotyping of 24 HPV types which included 15 high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82), 3 putative high-risk types (26, 53, and 66), and 6 low-risk types (6, 11, 42, 43, 44, and 70) was carried out in 270 cervical biopsies using Multiplex HPV Genotyping assay based on Luminex xMAP technology. As per the manufacturer’s instructions, PCR was carried out using sets of biotinylated broad range primers in a total volume of 50 µL containing 3.5 mM MgCl₂, 200 µM deoxynucleotide triphosphates, 0.75 unit of Taq DNA polymerase and 1µl primer mix. The amplification steps included an initial DNA denaturation at 94°C for 5 min, followed by 40 cycles of denaturation for 20 s at 94°C, annealing for 30 s at 38°C, and extension for 1 min 20 s at 71°C, before a final
extension for 4 min at 71°C. PCR positive samples were then subjected to the Luminex run. Ten microlitres of the PCR product was mixed with the Luminex bead mix containing distinct bead populations coupled to 24 HPV types. After thermal denaturation the target sequences were hybridized to bead-bound probes. The hybridized PCR products were labeled by binding to R-phycoerythrin conjugated streptavidin. The read-out was obtained in the Luminex bioanalyzer. HPV types were discerned according to the unique bead signature, whereas the presence of PCR products was determined by phycoerythrin fluorescence. An analytical sensitivity cut-off was calculated based on the negative control which was deducted from each of the read-out.

**HPV genotyping by PCR using MY09/11 and SPF1/2 primers:** The samples that tested HPV negative by luminex array (n=92) were further screened for HPV by PCR using MY09/11 L1 (Gravitt et al., 2000) and SPF1/2 (Kleter et al., 1998) primers.

**Association of HPV16, HPV18 and HPV16/18 infection with clinical outcome:** The genotyping data for the two HR-HPV types, HPV16, HPV18 and HPV16/18 together, where adequate follow-up data was available was compared with the clinical outcome of the patients. Kaplan-Meier analysis was done to determine association between infection with these HR-HPV types and recurrence of disease.

**Identification of integration site by APOT assay:** Site of viral integration was identified in a subset of cervical tumour samples (n=86) positive for HPV16, HPV18 or both and with a follow up data of 3 years using Amplification of Papillomavirus Oncogene Transcripts (APOT) assay as described by Klaes et al (Klaes et al., 1999). The assay is based on a 3’-rapid amplification of cDNA ends PCR and involves the following steps-

**Reverse transcription**-1µg RNA was reverse transcribed using SuperscriptTM first strand synthesis and an oligo (dT)17 primer coupled to a linker sequence referred to as
Frohman primer (Frohman et al., 1988). Integrity of the cDNA was checked by β-actin PCR.

 Nested PCR- For the PCR reactions HPV E7 primers was used as forward primers, an adapter primer complementary to the linker sequence in the Frohman primer as first reverse primer, and the Frohman primer as the second nested reverse primer. The PCR amplification was carried out in a total volume of 50µL containing 2.5mM MgCl₂, 10mM deoxynucleotide triphosphates, 25µM of each primer and 1 unit of Taq DNA polymerase. The PCR comprised of an initial DNA denaturation step of 94°C for 2min, followed by 35 cycles of denaturation for 30s at 94°C, annealing for 30s at 58°C, and extension for 4min at 72°C. A final extension for 20min at 72°C was given. 7µl of the PCR product was used as template for nested PCR. The PCR conditions were same as that of first PCR except the annealing temperature was 66°C.

 Cloning and sequence analysis- The amplified products thus obtained were visualized by 1.2% agarose gel electrophoresis. Amplicons other than the major episomal transcripts (~1050 bp for HPV 16 and ~1000 bp for HPV 18) were suspected to be derived from the integrated HPV genomes. These were excised from the gel and DNA isolated using GFX PCR DNA and Gel Band Purification Kit. The isolated DNA was either sequenced directly or after cloning into pTZ57R/T vector using the InsTA PCR Cloning Kit on the DNA automated sequencer. The chromosomal integration loci were determined using National Centre for Biotechnology Information (BLAST) and/or the University of California, Santa Cruz (BLAT) hg19 (Feb. 2009) human genome assemblies. Further the integration sites were checked for the presence of fragile sites and any genes of known identity by using NCBI fragile site map viewer and the UCSC Blat tool respectively.
**Validation of the recurrent integration sites:** Some of the recurrent integration sites were checked at the genomic level by carrying out genomic DNA PCR with HPV E7 primers as the forward primer and primers specific to a given chromosomal region as the reverse one.

**Association of viral integration with clinical outcome:** The data obtained was compared with the clinical outcome of the patients. Kaplan-Meier analysis was done to determine the association of the viral state (episomal/integrated) with recurrence of the disease. Disease free survival was considered from start of radiation therapy to the time when recurrence occurred or till last follow-up. Statistical significance was evaluated using the log-rank test. Further, where viral integration was observed at a particular chromosomal locus in ≥4 cases, Kaplan-Meier analysis was performed to check whether the site of integration in the host genome had any bearing on the disease prognosis.

**Detection of HPV copy number by quantitative real time PCR (qRT-PCR):** Copy number of two high-risk HPV types- HPV16 and HPV18 was assessed in the 86 samples where the physical status of the virus was already known by SYBR green based qRT-PCR. A 100 bp sequence located in the E7 gene of HPV16 or HPV18 was used for amplification, while a DNA fragment in the Tata Binding protein (TBP) gene which is a single copy gene was used as reference for relative quantitation. PCR conditions were 15 min at 95 ºC and 45 cycles (15 s at 95ºC, 30 s at 62ºC, and 1 min at 72ºC). HPV copy number was estimated by the $2^{-\Delta\text{CT}}$, with the TBP sequence as a reference for two DNA copies as described by Peter et al. (Peter et al., 2010).

**Identification of genetic alterations by Exome Sequencing:** Exome sequencing of six cervical tumour tissues (all of stage IIIB and with more than 70% tumour) and three matched controls (respective blood sample from three patients) was carried out at
Genotypic Technology (P) Ltd, Bangalore. Following steps were involved in the process:-

**Paired End library preparation**- This included shearing of the genomic DNA to less than 800bp fragments, ligation of adapters, gel extraction of ligated products and finally amplification of adapter-ligated library.

**Hybridization**- Hybridization of the DNA library to biotinylated RNA library was carried out, followed by selection of DNA-RNA hybrids by streptavidin coated magnetic beads and digestion of RNA.

**Post-Hyb processing**- This included PCR amplification, purification and finally sequencing of the captured library on Illumina GAIIX sequencing platform.

**Analysis for identification of single nucleotide variations and indels**:- The first level of analysis involved filtration of the reads by the software SeqQC. The filtered reads were then aligned against the *Homo sapiens* genome (Hg19 UCSC Build) by the program BWA 0.5.7 followed by variant calling from the alignment by Samtools 0.7.1a. In order to further shortlist the potential somatic variations (mutations) that could be important in cervical carcinogenesis, data from only the 3 paired samples was considered. The variations were first filtered out for the somatic and nonsynonymous ones and then subjected to Gene Ontology (GO) classification and KEGG pathway analysis to cluster the genes harboring the variations into groups based on their molecular function and biological processes. Important members were selected from this based on their relevance to cancer and the variations within them were subjected to prediction tools such as SIFT (Kumar et al., 2009) and Polyphen 2 (Adzhubei et al., 2010) for identifying whether a given amino acid substitution affects protein function.

**Validation of the data by Sanger sequencing**: Seventy one variations that included SNVs as well as short indels, in 69 genes were selected for validation by PCR followed
by sequencing. These variations were selected mainly based on their molecular function and biological processes using Gene Ontology (GO) classification and KEGG pathway analysis. The SIFT and Polyphen score were also considered.

**Validation of the data by Customized SNP Array:** Apart from validation by Sanger sequencing, we attempted to validate the exome sequencing data by customized SNP array. Accordingly, 700 known (reported in dbSNP version131) and 3178 novel nonsynonymous variations across the samples were selected for validation on an 8x60K customized Agilent oligonucleotide microarray. Validation of the few known variations was done as a check for the quality of the exome sequencing data.

**Analysis for detection of copy number variation (CNV) and loss of heterozygosity (LOH):** Identification of LOH and CNV was done from the exome sequencing only for the paired sample where the coverage between blood and tumour was comparable using ‘ExomeCNV’ package as described by Sathirapongsasuti et al. (Sathirapongsasuti et al., 2011).

**RESULTS**

**High-throughput HPV Genotyping by Luminex bead-based array:** Using GP5^+/GP6^+ primer set provided in the multiplex HPV genotyping kit:

- 178/270 samples were found to be positive for HPV.
- 169/178 samples were positive for different HPV types whereas 9 samples were negative. These 9 samples could have HPV infection not included in the 24 types detected by the kit.
- Infection with HPV16 and/or HPV18 infection was most common - 114 samples being positive for HPV16 alone, 6 samples for HPV18 alone and 16 samples for both HPV16 and HPV18.
Frequently reported high-risk HPV type 31, 33 and 35 infection was found to be low.

**HPV genotyping by PCR using MY09/11 and SPF1/2 primers:** In order to estimate the true HPV positivity in the 270 cases, the 92 cervical cancer biopsies negative for HPV by luminex array, were subjected to PCR using MY09/11 and SPF1/2 primers.

- 25/92 samples were positive for HPV by MY09/11 PCR.
- 79/92 samples were positive by SPF1/2 primers. These 79 samples also included the 25 samples that tested HPV positive by MY09/11 PCR.
- The overall HPV positivity was therefore 95% (257/270).
- Further genotyping of the 79 samples, using HPV 16/18 specific primers, showed 49 samples to be positive for HPV16.
- Overall HPV16/18 positivity in this cohort was 69% (185/257).

**Association with clinical outcome:** Kaplan-Meier survival analysis data for 125 patients with HPV type16, 18 and dual infection and with adequate clinical follow-up revealed that there was no significant difference between infection with these two HR-HPV types in terms of disease outcome.

**Identification of viral integration sites by APOT assay:** In order to detect physical state and/or site of integration of the virus in 86 cases, APOT assay was performed. Sequencing data revealed that –

- 79% (68/86) of the cervical cancer samples showed HPV integration
- Rest 21% (18/86) had HPV in the episomal form
- In 18% of the patients (n=12) integrated as well as episomal form of HPV was observed
- The site of integration could be predicted with a high score in 48 cases, for the remaining 20 cases the score was low. Only those cases where the integration site was
predicted with high score (n=48) were analyzed further for different features associated with the same.

- Only 1 sample showed HPV integration at two chromosomal loci simultaneously
- Using NCBI Fragile site Map Viewer it was observed that 60% of integrations (29/48) were located in or close (~5Mb) to a common or rare fragile site
- Using the UCSC Blat tool 58% of the sequences (28/48) were seen to have homology with protein coding genes. These genes belonged to various categories ranging from oncogenes, transcription factors, and tumour suppressor genes

**Validation of the recurrent integration sites:** For most of recurrent integration sites such 1p36.23, 3q28, 3q23, 6q22.31, 6q23.3, 8p11.21 11q13.1 and 13q22.1 PCR amplification was observed with genomic DNA, confirming genomic integration.

**Association of viral integration with clinical outcome:** Survival data revealed that 16 out of 18 patients with only episomal form of HPV (16/18), had disease free survival as compared to those with integrated form of the virus, indicating a good clinical outcome (p=0.067, representing a borderline significance). Further, in case of viral integration, patients with integration at chromosomal loci 1p (7/7), 6q (4/4) and 11q (4/4) were disease free, while most of the patients with integration at the chromosomal loci 3q (5/8), 13q (4/4) and 20q (2/4) showed recurrence of the disease in the form of either loco-regional or distant metastasis.

**Detection of HPV copy number by quantitative real time PCR (qRT-PCR):** Viral copy number as high as 443 to as low as 1 was detected across the 86 samples with known viral physical status. Comparison with survival of the patients as well as physical status revealed that cases where the virus was in the episomal form, the survival had no association with copy number. But those with integrated form of the virus showed reduced survival when the copy number was high as compared to when it was low.
Therefore, copy number of the virus combined with the physical state might serve as a good prognostic marker for the disease.

**Exome capture, sequencing and analysis:** After variant calling, about 14400 unique sequence variants including single nucleotide variations and indels across all 9 samples were detected. Of all these variations, a total of 7407 SNVs, 94 small deletions (1-5bp) and 71 small insertions (1-3 bp) were not reported in dbSNP131 and were represented as putative novel sequence variants. These variations could be unique to the Indian population. Further analysis revealed that about 6114 of the SNVs to be nonsynonymous. Analysis with the paired data identified 874 novel, potential somatic SNVs, 30 small deletions (1-5 bp) and 22 small insertions (1-3 bp). Out of these, 395 SNVs, 12 small deletions and 5 small insertions were found to have defined/ important functions. SIFT analysis predicted 167 variations to have a potential deleterious effect on the protein function while with Polyphen 2, 128 variations were predicted to be damaging. Both the prediction tools predicted 73 common variations to be deleterious.

**Validation of the data by Sanger sequencing:** So far, we have tried validating 71 different variations in 69 genes by PCR followed by sequencing. Most of the variations did not get validated by this method. This may be due to the fact that Sanger can accurately detect 1:2 or 1:3 representations of heterozygotes and since almost all of these variations were heterozygous, there is a chance of missing out the mutant allele. However, certain variations in genes like- RNASEL (R462Q), PTPRJ (R1222H), ZFP64 (Y210X), NFAT5 (Y1188C) and FBN1 (delAGG), were validated by sanger sequencing. But when we checked for these variations in the blood, they were detected in the blood as well. Therefore, these cannot be considered as somatic as given by exome sequencing data. These could however represent variations that might predispose women to cervical cancer.
**Synopsis**

*Validation of the data by Customized SNP Array:* In the SNP array 1455/3178 novel nonsynonymous and 192/700 known single nucleotide variations were validated. The overall concordance rate between the exome sequencing and SNP array data was found to be 33%. However, non confirmation in the array did not mean that a particular variation call was incorrect since a very high threshold was used as cut off while analyzing the array data. Although, most of the SNVs that were validated were germline in nature, a set of 22 novel, nonsynonymous and somatic variations were obtained which included important candidates such as CD97, YWHAH and CD52. These somatic SNVs might help in understanding the mutational landscape of cervical cancer.

*CNV and LOH identification:* CNV in chromosomes 3, 8 and 17 and LOH in chromosomes 2, 3, 5 were observed for one of the paired samples. Though this data has not been validated, this is in agreement with previous published reports.

**DISCUSSION**

We have tried to obtain a broad overview of cervical cancer by addressing various factors – type of HPV infection, viral copy number, viral integration and genetic alterations that might play a role in cervical carcinogenesis in a cohort of Indian women with locally advanced cervical cancer.

The HPV genotyping study demonstrated 95% HPV positivity using three different primer sets. It is apparent from the results that a single set of primers is not sufficient to estimate the true HPV infectivity. Of the various HPV genotypes HPV16 was most common (60%), followed by infection with HPV18 alone (2%). Dual infection with HPV16/18 (6%) and HPV16/45 (3%) was also observed. These results are in concordance with other studies from the Indian subcontinent which reports 57-
65% HPV16 positivity, followed by HPV18, 45, and 33 in cervical neoplasia (Basu et al., 2009; Bhatla et al., 2008).

For the viral integration study, we chose APOT assay in order to limit the analysis to integration sites with a transcriptionally active viral genome. Also, APOT assay allows detection of integrated viral genome in clinical lesions even in the presence of a large excess of nonintegrated episomal form of viral genomes (Klaes et al., 1999; Vinokurova et al., 2008). The frequency of viral integration into the host genome in cervical carcinomas has been reported to be as high as 100% in HPV18+ tumours (Corden et al. 1999) and up to 80% in HPV16+ tumours (Melsheimer et al., 2004). In our study we found 4 HPV18+ samples where the virus was integrated and one HPV18+ sample where the virus was episomal. The incidence of integration in HPV16+ samples was 94%. The preference of integration for the chromosomal loci 1p, 3q, 6q, 11q, 13q, 6q and 20q indicate that integration might not be a random event. Most of the integrations (28/48) were found to be located within or near certain genes. This could indicate that the virus prefers transcriptionally active regions for the integration event. Such genes included oncogenes such as myc, transcription factors like TP63, MECOM, etc. The importance of genes like myc and TP63 has already been demonstrated in the context of HPV integration (Wentzensen et al., 2002). Again, integration of the virus near or within fragile sites is a frequently reported phenomenon (Kraus et al., 2008; Thorland et al., 2000). In our study 29/48 integrations were located within or near a common or rare fragile site which is in concordance with previous reports. Comparison of the physical state of the virus (episomal/integrated) with the clinical outcome after radical radiotherapy revealed that patients with episomal form of the virus had increased disease free survival compared to those with integrated form. This observation is supported by various reports which state that the integration event is associated with a
decreased disease free survival (Vernon et al., 1997). However, there are contrasting reports as well, according to which physical state of the virus does not correlate with disease free survival. This needs to be studied further.

Though infection by HPV plays a major role in the etiology of cervical cancer, possible role of several genetic alterations, such as Point mutations, Chromosomal aberrations etc. cannot be completely ruled out. Especially among older women genetic mutations may be a reason for cervical cancer acting in synergy with a lower immune response.

Studies have identified genes such as LKB1, FGFR3, PI3KA, Ras, Smad, etc. to be frequently mutated in cancer of the cervix (Cui et al., 2009; Maliekal et al., 2003; Wong et al., 1995; Yee et al., 2000). However, these studies were mostly based on candidate gene approach and till date there are no reports describing the mutational landscape of the disease. Therefore, we attempted to identify the alterations that might play role in cervical carcinogenesis by carrying out whole-exome sequencing of 6 locally advanced cervical biopsies (3 paired and 3 unpaired). Apart from helping understanding the mutational landscape of cervical cancer, these novel variations would be a valuable source of information about the Indian genome which the dbSNP currently lacks. We obtained about 6114 novel nonsynonymous variations unique to this population. A proportion of this data was also validated by customized DNA microarray and sanger sequencing. The significance of the exome sequencing analysis lies in the fact that it provides significant information about the Indian genome with special reference to cervical cancer. To the best of our knowledge, this is the first report where we have applied whole exome sequencing for identifying single nucleotide variations, short indels, LOH and CNV in cervical cancer in Indian women.
REFERENCES


