## 3.1 Materials

### 3.1.1 Survey for cancer prevalence

3.1.1.1 Survey of International Cancer Registries

Gender-wise incidence and mortality rates, age-wise distribution in the two age-groups of $< 65$ and $\geq 65$, and 5 years prevalence of NPC were collected from the 2012 global cancer databases (GLOBOCAN) which is maintained by the International agency for research on cancer (IARC) of the World Health Organization (WHO).

3.1.1.2 Survey of National Cancer Registries

Incidence data of NPC (2009-2012) were obtained from Population Based Cancer Registries (PBCRs) and Hospital Based Cancer Registries (HBCRs) database, and the Atlas of Cancer in India databases under the National Cancer Registry Programme (NCRP) of Indian Council of Medical Research.

### 3.1.2 Study design

The present study was an age, sex and ethnicity matched population based case, first degree relative and control study. The study duration was from May 2012-April 2015. The work was carried out at Biotechnology Department, Assam University, Silchar. Incident cases, first degree relatives (FDRs) and controls subjects willing to participate in the study were indigenous people of Nagaland, Mizoram and Manipur. Demographic information was recorded in a structured pre-designed questionnaire. All the subjects were interviewed regarding tobacco habits (smoking and betel quid chewing), alcohol drinking and smoked meat and fermented fish intake. For dietary habits, subjects were divided to never (who do not consume), regularly (who consumed weekly or more) and occasionally (consuming monthly or biweekly). Smokers and chewers were defined as having smoked or chewed at least 1/day for six months. Those

who had not smoked or chewed betel quid were defined as non smokers and chewers. They were further categorized based on their frequency of consumption. Heavy smokers were those who smoked at least ≥ 20 bidis/cigarettes per day for ≥ 20 years, and light smokers where those who smoked < 20 bidis/cigarettes per day for < 20 years. Whereas, heavy chewers where those who chewed ≥ 10 doses per day for ≥ 20 years, and those who chewed < 10 doses per day for < 20 years were defined as light chewers. Similarly, subjects who had drunk alcoholic beverages at least once a week for more than one year previously were defined as drinkers, and non-drinkers were those who had not drunk alcohol. Alcohol drinking was categorized as light drinkers (< 5 drinks per week for < 20 years) and heavy drinkers (≥ 5 drinks per week for ≥ 20 years). The patients were divided into two age groups; a younger group of below or 50 years of age and an older group of above 50 years. A total of 123 NPC patients, 90 FDRs and 189 controls were included in the study.

3.1.2.1 Ethical Clearance

Ethical clearance **(IEC/AUS/2013-011dt-20/3/13)** for the present study was taken from the Institutional Review board (IRB) and written consent was taken from individual patients or relatives involved in the study.

3.1.2.2 Selection of subjects (Cases/FDRs/Controls)

Cases were identified through the population based cancer registries (PBCRs) in the Civil Hospital, Aizwal; Regional Institute of Medical Sciences, Imphal and Naga Hospital Administration, Kohima. The inclusion criteria for the cases includes histopathologically confirmed NPC cases with no evidence of pulmonary inflammation or benign NPC tumours and were resident of Nagaland, Manipur and Mizoram. The exclusion criteria for the cases were patient's diagnosis reported by death certificate or

at autopsy, diagnosed with oral, throat, stomach or lung cancers, cases too old or too debilitated to be interviewed elaborately and those who refuse to participate in the study. The inclusion criteria for first degree relatives were those who had first degree blood relation with NPC cases, who can be either death or alive. The inclusion criteria for controls were those who had no blood relation with NPC cases, non malignant individual, residence of Nagaland, Manipur and Mizoram, age, sex and ethnicity matched with cases.

3.1.2.3 Collection of samples

Sample of either blood or buccal swab or both were collected from each subject included in the study. 5 ml of venous blood were collected using sterilized needles and stored in EDTA vials. The buccal swabs were collected by using cotton buds and were stored in eppendorf tubes containing 500µl of TES buffer. Formalin fixed paraffin embedded tissues (FFPE) of NPC patients were also obtained from the participating hospitals.

### 3.1.3 Chemicals and Reagents

The chemical used were mostly of molecular grade and includes the following:

a) Tris base (hydroxymethyl amino methane), Tris acetate
b) Concentrated HCl
c) Ethylene diamine tetrachloro acetic acid (EDTA)
d) Sodium chloride (NaCl)
e) Sodium lauryl sulfate /Sodium dodecyl sulphate (SDS)
f) Proteinase K (Promega USA)
g) Phenol:Chlorofom:Isomaylacohol (25:24:1)
h) Chloroform:Isoamyalcohol (24:1)
i) Sodium hydroxide pellets
j) Absolute ethanol (Bengal Chemical and Pharmaceutical Works Ltd, Kolkata.)

k) Ethidium bromide
l) Bromophenol blue
m) Sucrose
n) Agarose (Ultra-pure)

3.1.3.1 Stock solution preparation

a) Tris (hydroxymethyl) aminomethane-HCl or Tris-HCl (1m, pH-8.0)

121.4 g of Tris base was dissolved in 500ml of distilled water. The pH was adjusted to 8 by adding concentrated HCl and the final volume was make upto 1 litre with distilled water. Autoclaved at 15 lbs for 10 minutes and stored at room temperature.

b) Ethylene diamine tetrachloro acetic acid or EDTA (0.5M, pH-8.0)

146.1 g of EDTA was dissolved in 500 ml of distilled water. The pH was adjusted to 8 by adding sodium hydroxide and the final volume was make upto 1 litre with distilled water. Autoclaved at 15 lbs for 10 minutes and stored at room temperature.

c) Sodium chloride or NaCl (5M)

292.2 g of NaCl was dissolved in 800 ml of distilled water and dissolved properly. The final volume was make upto 1 litre and stored at room temperature.

d) Sodium Dodecyl Sulfate or SDS (10%)

10 g of SDS was dissolved in 100 ml of distilled water and stored at room temperature.

e) Ethidium bromide or ETBr (10 mg/ml)

10mg of ethidium bromide was dissolved in 1 ml of nuclease free water. Mixed properly and stored at room temperature after wrapping in aluminium foil.

f) Proteinase K

The enzyme is dissolved in nuclease free water to a final concentration of 20 µg/ml.

3.1.3.2 Buffers preparation

a) TES buffer (Tris-HCl EDTA Sodium chloride) – 10 ml

| | |
|---|---|
| Tris-HCl | 50 mM (added 500 µl from stock 1 M Tris-HCl) |
| EDTA | 25 mM (added 500 µl from stock 0.5 M EDTA) |
| NaCl | 150 mM (added 300 µl from stock 5 M NaCl) |
| Distilled water | 8.7 ml |

b) Lysis buffer – 10 ml

| | |
|---|---|
| Tris-HCl | 10 mM (added 100 µl from stock 1 M Tris-HCl) |
| EDTA | 10 mM (added 250 µl from stock 0.5 M EDTA) |
| NaCl | 50 mM (added 100 µl from stock 5 M NaCl) |
| SDS | 20% (100 µl) |
| Distilled water | 8.6 ml |

c) TAE buffer – 10 ml (50X)

The buffer has the following components:

| | |
|---|---|
| Tris base | 242 g |
| Glacial acetic acid | 57.1 ml |
| EDTA (0.5 M) | 100 ml |

TAE buffer is commonly prepared as a 50X stock solution for laboratory use. A 50X stock solution can be prepared by dissolving 242 g Tris base in water,

adding 57.1 ml glacial acetic acid, and 100 ml of 0.5 M EDTA (pH 8.0) solution and bringing the final volume up to 1 litre. This stock solution can be diluted 49:1 with water to make a 1X working solution. This 1X solution will contain 40mM Tris, 20mM acetic acid, and 1mM EDTA.

d) TE buffer (1X)- 100ml

A typical recipe for making 1X TE buffer is:

10 mM Tris, bring to pH 8.0 with HCl

1 mM EDTA

To make a 100 ml solution of TE Buffer, 1 ml of 1 M Tris-HCl (pH 8.0) and 0.2 ml EDTA (0.5 M) and made up with double distilled water up to 100ml.

e) 6X loading dye

The buffer has the following composition:

| | |
|---|---|
| Sucrose | 40% (w/v) |
| Bromophenol blue | 0.25% |

The loading buffer was prepared by dissolving 4 g of sucrose and 25 mg of bromophenol blue in 10 ml of distilled water and stored at 4 $^{\circ}$C to avoid mould growing in the sucrose.

**3.2 Methods**

3.2.1 Isolation of genomic DNA

DNA was isolated from tissues as well as from peripheral blood by following standard Phenol: Chloroform extraction protocol (Ghosh and Mondal, 2012). The isolated DNA was stored in TE buffer at -86$^{o}$C for further use.

3.2.1.1 Isolation from buccal swap

1. Buccal swabs were collected using cotton buds in a sterilized microcentrifuge tube of 1.5 ml containing 500 µl of TES buffer.
2. Cotton buds were carefully dissolved in the microcentrifuge tubes and squeezed assuming all the buccal swabs cells retain in the TES buffer and is removed.
3. To the above sample, 5-8 µl of Proteinase K and 80 µl of 10% SDS
4. Mixed the samples thoroughly and then incubated at 56ºC for three hours in water bath.
5. The samples were taken out of the water bath and equal amount of Phenol: Chloroform: Isoamyl alcohol (500 µl in 25:24:1) were carefully added.
6. Closed the tubes tightly and vortex it briefly.
7. The tubes were then centrifuged at 13000 rpm for 10 minutes.
8. Carefully removed the aqueous layer for each sample into new clean micro-centrifuge tubes. Then added equal volume of Chloroform: isoamyl alcohol (24:1) into the micro-centrifuge tubes and vortex briefly for uniform mixing.
9. Again the tubes were centrifuged at 13000 rpm for 10 minutes.
10. Transfer the aqueous layer for each sample into new, clean micro-centrifuge tubes. Then added double the volume of 100% chilled ethanol to the tubes.
11. The tubes were inverted to mix and precipitated the DNA.
12. Kept the samples in - 20° C for overnight incubation for precipitation or 86° C for 15-30 mins.
13. Centrifuged the tubes at 13,000 rpm for 10 minutes.
14. The supernatant was discarded and the pellet was suspended in 500 µl 70% ethanol. Again centrifuged the tubes at 13,000 rpm for 10 minutes.
15. The supernatant was discarded and the pellet was air dried under laminar air flow and re-suspended in 20-30 µl of Nuclease free water or 1X TE buffer.
16. The tubes were kept at 56ºC for 15 minutes and stored at -20 ºC for further use or 86° C for long term preservation

3.2.1.2 Isolation from whole blood

1. Blood samples typically were obtained as 2 ml of whole blood in EDTA vacutainer tubes frozen at 4 °C.
2. The blood samples of 80-100 µl was pipetted in a sterilized microcentrifuge tube of 1.5 ml containing 500 µl of lysis buffer
3. To the above sample, 2-4 µl of Proteinase K and 50µl of 10% SDS
4. Mixed the samples thoroughly and then incubated at 56ºC for three hours in water bath.
5. The samples were taken out of the water bath and equal amount of Phenol: Chloroform: Isoamyl alcohol (500 µl in 25:24:1) were carefully added.
6. Closed the tubes tightly and vortex it briefly.
7. The tubes were then centrifuged at 13000 rpm for 10 minutes.
8. Carefully removed the aqueous layer for each sample into new clean micro-centrifuge tubes. Then added equal volume of Chloroform: isoamyl alcohol (24:1) into the micro-centrifuge tubes and vortex briefly for uniform mixing.
9. Again the tubes were centrifuged at 13000 rpm for 10 minutes.
10. Transfer the aqueous layer for each sample into new, clean micro-centrifuge tubes. Then added double the volume of 100% chilled ethanol to the tubes.
11. The tubes were inverted to mix and precipitated the DNA.
12. Kept the samples in - 20° C for overnight incubation for precipitation or 86° C for 15-30 mins.
13. Centrifuged the tubes at 13,000 rpm for 10 minutes.
14. The supernatant was discarded and the pellet was suspended in 500 µl 70% ethanol. Again centrifuged the tubes at 13,000 rpm for 10 minutes.
15. The supernatant was discarded and the pellet was air dried under laminar air flow and re-suspended in 20-30 µl of Nuclease free water or 1X TE buffer.
16. The tubes were kept at 56ºC for 15 minutes and stored at -20 ºC for further use or 86° C for long term preservation.

3.2.2 Quality and quantity analysis of the extracted DNA

The isolated DNA must be free from most of the associated proteins that keeps DNA coiled and should be of adequate quantity. Therefore, the quality and quantity of the extracted DNA should be checked before undergoing downstream process like PCR.

3.2.2.1 Spectrophotometric Analysis

The quality and quantity of the extracted DNA was determined spectrophotometrically as follows:

1. 50 µl of nuclease free water was taken in a cuvette and the spectrophotometer was calibrated at 260 nm and 280 nm.
2. The isolated DNA stock solution was then taken at different dilution rate viz. 200, 50 and 10 times by adding nuclease free water.
3. 2 µl of each DNA samples along with 48 µl of nuclease free water were added in a cuvette and mixed properly.
4. Optical densities (OD) were measured at 260 nm ($OD_{260}$) and 280 nm ($OD_{280}$) in an UV spectrophotometer (Biophotometer, Eppendrof) against nuclease free water as blank. The yield and purity of DNA samples were estimated as follows:

Concentration of DNA stock solution (µg/ml) = $OD_{260}$ X 50 (dilution factor) µg/ml. Purity of DNA stock solution = $OD_{260}$/ $OD_{280}$ (for pure DNA samples this ratio must be in the range of 1.80 – 2.00). From the concentration of DNA stock solution, the total yield of DNA was calculated and recorded.

3.2.2.2 Agarose gel electrophoresis for genomic DNA quality and quantity analysis

Agarose gel concentration of 0.8 % (0.8 µg/ml) was prepared for the analysis of the isolated genomic DNA samples. Ethidium bromide (10 mg/ml) was added to the gel to facilitate the visualization of the DNA after electrophoresis. Samples containing DNA mixed with loading buffer (bromophenol blue) were then pipetted into the sample wells alone side with known molecular weight DNA markers and leave for electrophoresis. To visualize the DNA, the gel was placed on an UV transilluminator. The quality and quantity of the isolated DNA was determined by comparing it with the molecular weight markers which give a rough idea regarding the concentration.

3.2.3 Polymerase chain reaction (PCR)

PCR was performed to amplify the genes and determined their respective genotypes. Multiplex PCR was performed to genotype *GSTT1* and *GSTM1* genes, PCR-RFLP for genotyping the *CY1A1 T3801C, XRCC1 Arg399Gln, XRCC1 Arg188His*, Sanger sequencing for validation of PCR-RFLP results and genotyping TNFRSF19, *CDK2BAS1*, and *MECOM* genes. The genotypes of *TNFRSF19, CDKN2B-AS1*, and *MECOM* genes as determined by Sanger sequencing were further validated by SNapShot multiplexed single-base extension sequencing. The following components are included in a PCR reaction setting. Each 20 µl PCR reaction mixture contains:

| | | |
|---|---|---|
| 10X PCR Buffer | : | 2µl |
| dNTPs (10mM) | : | 2µl |
| Forward primers (20pmoles/µl) | : | 0.5µl |
| Reverse primers (20pmoles/µl) | : | 0.5µl |
| High fidelity DNA polymerase (5U/µl) | : | 0.4µl |
| Genomic DNA (100-200ng) | : | variable |
| Nuclease free water | : | upto 20µl |

3.2.3.1 Primer designing and optimisation

PCR primers used in this study were newly designed, or were published primers. Primers were designed and validated using NCBI /Primer-BLAST tools (http://www.ncbi.nlm.nih.gov/tools/primerblast/). PCR conditions for each primer (Table 3.1) were optimized.

**Table 3.1** List of primers use for amplification of different genes

| Primer name | Primer sequence | Gene/Loci | Product size |
|---|---|---|---|
| *GSTT1*-F<br>*GSTT1*-R | 5′-TTCCTTACTGGTCCTCACATTCTC-3′<br>5/-TCACGGGATCATGGCCAGCA-3/ | *GSTT1* | 480bp |
| *GSTM1*-F<br>*GSTM1*-R | 5/-GAACTCCCTGAAAAGCTAAAGC-3/<br>5/-GTTGGGCTCAAATATACGGTGG-3/ | *GSTM1* | 215bp |
| *CYP1A1*-F*<br>*CYP1A1*-R* | 5'-GAACTGCCACTTCAGCTGTCT-3/<br>5/-GCTGCATTTGGAAGTGCTC-3/ | *CYP1A1* | 312bp |
| *CYP1A1*-F<br>*CYP1A1*-R | 5/-TAGGAGTCTTGTCTCATGCCTT-3<br>5/-CAGTGAAGAGGTGTAGCCGCT-3/ | *CYP1A1*<br>(T3801C) | 343bp |
| *LMP1*-F<br>*LMP1*-R | 5′-GGAGCCCTTTGTCTACTCCTACT-3/<br>5′-TGCCTGTCCGTGAAATTC-3/ | *EBV-LMP1* | 150bp |
| *XRCC1*-F<br>*XRCC1*-R | 5′-TTGTGCTTTCTCTGTGTCCA-3/<br>5′-TCCTCCAGCCTTTTCTGATA-3/ | *XRCC1*<br>(Arg399Gln) | 615bp |
| *XRCC2*-F<br>*XRCC2*-R | 5′-TCACCCATCTCTCTGCCTTTTG-3/<br>5′-TTCTGATGAGCTCGAGGCTTTC-3/ | *XRCC2*<br>(Arg188His) | 205bp |
| *TNF19*-F<br>*TNF19*-R | 5/-TGTAAGACAAGCCACGTGATTC-3/<br>5/-TAATGAGCAAATACGTCTCAAAG-3/ | *TNFRSF19*<br>(rs9510787) | 268bp |
| *CDK*-F<br>*CDK*-R | 5/-CTTTCCTGCCCTTTTGCCTCATA-3/<br>5/-GTTGTGAAAGCAGTTGTTTACG-3/ | *CDKN2BAS1*<br>(rs1412829) | 502bp |
| *MECOM*-F<br>*MECOM*-R | 5/-GTCACTATCCAAAAGCAGTTGCC-3/<br>5/-TTCTTGCCCAATGGATGATACTC-3/ | *MECOM*<br>(rs6774494) | 387bp |
| *C tract*-F<br>*C tract*-R | 5/-CAGGGTCATAAAGCCTAAATAG-3/<br>5/- GAGGTAAGCTACATAAACTGTG -3/ | mt D-loop | 109bp |
| *GAPDH*-F<br>*GAPDH*-R | 5/- GAAATCCCATCACCATCTTCC-3/<br>5/- GAGCCCCAGCCTTCTCCATG-3/ | *GAPDH* | 105bp |

** Exon 7 of CYP1A1 gene as an internal control in multiplex PCR*

3.2.3.2 Restriction enzyme digestion

Restriction enzymes are isolated from bacteria which have the ability to recognize specific DNA sequences and cut within it at particular specific site called restriction site. Isolated restriction enzymes are used to manipulate DNA for different scientific applications. For examples, gene cloning, allelic differences (Single nucleotide Polymorhism SNP), mutation detection etc. In this study *MspI*, *HpaII* and *HphI* (New England BioLabs, USA) restriction enzymes were used to detect allelic differences. Material required for restriction digestion includes 10x buffer, restriction enzyme, nuclease free water, 1.5 ml Eppendorf tubes, pipette, and water-bath for incubation. Each 15 µl reaction of restriction enzyme contains:

| | | |
|---|---|---|
| 10X PCR Buffer | : | 1.5 µl |
| Restriction enzymes (20U/ µl) | : | 1 µl |
| PCR products | : | 7 µl |
| Nuclease free water | : | upto 15µl |

All the above contents were mixed in a 0.2ml autoclaved PCR tube and incubated at 37°C overnight.

3.2.3.3 Multiplex PCR genotyping for *GSTT1* and *GSTM1* metabolic genes

Genotyping of *GSTT1* and *GSTM1* genes were performed by multiplex PCR assays using exon 7 of *CYP1A1* gene as an internal control (Mondal et al., 2013). The reaction was conducted at 94°C for 5 min for the initial denaturation, following 30 cycles of denaturation at 94°C for 30 s, annealing at 59°C for 45 s, extension at 72°C for 30 s and final extension at 72°C for 5 mins. A 480, 315 and 215bp amplicons represents the *GSTT1*, *CYP1A1* and *GSTM1* genes.

3.2.3.4 PCR-RFLP genotyping of *CYP1A1* T3801C polymorphism

PCR amplification containing the *CYP1A1* T3801C polymorphic site was conducted at 94°C for 5 min for the initial denaturation, following 30 cycles of denaturation at 94°C for 30 s, annealing at 62°C for 45 s, extension at 72°C for 30 s and final extension at 72°C for 5 mins (Kiruthiga et al., 2011). For the *CYP1A1* T3801C genotype analysis, a 343bp PCR product was digested by the *MspI* restriction enzyme (New England BioLabs, USA); a single 343 bp fragment represents the wild-type allele (TT), three fragments of 343, 200 and 143 bp indicates for the heterozygous variant (TC) and two fragments of 200 and 143 bp for the homozygous variant allele (CC). The RFLP results were confirmed by sequencing of the PCR amplified product using Genetic Analyzer 3500, Applied Biosystems (Molecular Medicine Lab, Department of Biotechnology, Assam University, Silchar, India).

3.2.3.5 PCR amplification of viral *LMP1* gene to detection of EBV infection

EBV infection was detected through PCR amplification of the *LMP1* gene. The primer specific for the amplification of *LMP1* gene was designed using the primers blast software of NCBI. The PCR programme used for amplification was: initial denaturation step at 94°C for 5 mins; 40 cycles of denaturation at 94°C for 30s; annealing at 62°C for 30s, extension at 72°C for 45s and final extension at 72°C for 5mins. A 150 bp product in 2% agarose gel determines the presence of the EBV-*LMP1* gene.

3.2.3.6 PCR-RFLP genotyping of *XRCC1* Arg399Gln polymorphism

In the *XRCC1 Arg399Gln* genotype analysis, a 615 bp fragment amplified (Choudhury et al., 2014). The PCR programme used for amplification was: initial denaturation step at 94°C for 5 mins; 30 cycles of denaturation at 94°C for 30s;

annealing at 58°C for 30s, extension at 72°C for 45s and final extension at 72°C for 5mins. The PCR amplicon was digested with *HpaII* restriction enzyme (New England BioLabs, USA); two fragments of 240 and 375 bp represents the wild-type allele GG (Arg/Arg), three fragments of 615, 375 and 240 bp indicates for heterozygous GA (Arg/Gln) and a single 615 bp fragment for the variant allele AA (Gln/Gln).

3.2.3.7 PCR-RFLP genotyping of *XRCC2* Arg188His polymorphism

The *XRCC2 Arg188His* genotype was determined by amplifying a 205 bp fragment (Choudhury et al., 2014). The PCR programme used for amplification was: initial denaturation step at 94°C for 5 mins; 30 cycles of denaturation at 94°C for 30s; annealing at 60°C for 30s, extension at 72°C for 45s and final extension at 72°C for 5mins. The resulting fragment was digested with *HphI* restriction enzyme (New England BioLabs, USA); a single 205 bp fragment represents the wild-type allele GG (Arg/Arg), three fragments of 205, 137 and 68 bp indicates for the heterozygous GA (Arg/Gln) and two fragments of 137 and 68 bp for the variant allele AA (His/His). The RFLP results were confirmed by sequencing 10% of the randomly selected samples from both cases and controls by Sanger sequencing using Genetic Analyzer 3500, Applied BioSystems (Molecular Medicine Lab, Department of Biotechnology, Assam University, Silchar, India).

3.2.3.8 PCR Amplification of Hotspot Genes Related to NPC

The *TNFRSF19, CDKN2B-AS1* and *MECOM* genes from the extracted genomic DNA were amplified with the help of PCR reaction mix using our own designed primer. Large scale multiplex amplification of the three hotspot genes was carried out by PCR containing 10x PCR buffer, 10mM of dNTPs, 2.5 Units of High fidelity DNA polymerase (*Pfu* enzyme), 0.25mM of MgCl2, 1% DMSO and 10-20

pmol of each primers. The thermal cycling of the PCR was carried out at 94°C for 5 min for initial denaturation temperature, followed by 35 cycles of denaturation at 94°C for 45 sec, annealing at 64°C for 45 sec, elongation at 72°C for 1min. A final elongation step was maintained at 72°C for 8 min. The PCR amplified product was analyzed in 1.5% agarose gel in TAE buffer with Ethidium Bromide staining (10 mg/ml).

3.2.3.9 Purification of PCR products by Gel purification

Purification of PCR products is required before sequencing the genes. Purification was done by GeneJET Gel Extraction Kit (Thermo Scientific, K0692) as per manufacturer protocol. The process is given below:

1. All the gel slices containing the amplified fragments of *CYP1A1, XRCC1, XRCC2, TNFRSF19, CDKN2BAS1* and *MECOM* using a clean scalpel was excised.
2. Then 1:1 volume of binding buffer was added to the each slice (volume : weight) (For example, 200µl of binding buffer was added for every 200mg of agarose gel slice)
3. The gel mixture was incubated at 56°C until the gel slices were completely dissolved. Then 200µl of Absolute ethanol was added to the solution.
4. All the samples were briefly vortexed and then the solution was transferred to GeneJET purification columns.
5. These were then centrifuged at 14000g for 1 minute. The flow through was discarded and the columns were placed back into the same collection tube.
6. 200µl of Pre-wash buffer (diluted with ethanol) was added to the GeneJET purification columns.
7. These were then centrifuged for 1 minute at 14000g. The flow through was discarded and the column was placed back into the same collection tube.
8. 700µl of Wash buffer (diluted with ethanol) was added to the GeneJET purification columns and then centrifuged for 1 minute at 14000g. Next, the

flow through was discarded and the columns were placed back into the same collection tube. This step was carried out 2 times.

9. The empty GeneJET purification columns were centrifuged to an additional 1 minute at 14000g to completely remove the residual wash buffer.
10. These purification columns were then transferred into clean 1.5ml microcentrifuge tubes.
11. 10 µl of elution buffer was added to the centre of the each purification column membranes.
12. It was kept in room temperature for 5 minutes.
13. It was then centrifuged for 1 minute at 14000g.
14. The GeneJET purification columns were discarded and the purified PCR products were stored at -20$^0$C for further downstream processes.

3.2.3.10 Sanger Sequencing

Sanger Sequencing was performed for genotyping *MECOM, TNFRSF 19* and *CDK2BAS1* genes as well as to validate the PCR-RFLP results of *CYP1A1 T3801C, XRCC1 Arg399Gln* and *XRCC2 Arg188His* polymorphisms. The amplified products were purified and sequenced bi-directionally using automated DNA Sequencer (ABI 3500 Genetic Analyzer; Applied Biosystem, Inc. USA). The sequencing reaction was performed using Ready reaction premix (2.5X), BigDye Terminator v3.1 dye Ssequencing buffer (5X); 20 pmol of the primer (either forward or reverse), the purified PCR products as template (3-10ng) and nuclease free water upto 20 µl. The Chain termination reaction was carried out at 25 cycles of 1 min at 94°C, 5 sec at 50°C and 4 min at 60°C. The fragments were then purified by Sodium Acetate/EDTA/Ethanol method prior to run in the 3500 Genetic Analyzer. The Sodium Acetate 3M, pH-5.2 and 125mM EDTA were added maintaining 1/10th of the total volume of the product followed by the addition of double the volume of absolute ethanol and consecutive

wash with 70% ethanol. The sample were dried properly and resuspended in Hi-Di Formamide (Applied Biosystems) prior to load in the capillary.

3.2.3.11 Relative quantitative of mitochondrial DNA copy number using real-time PCR

Relative quantification describes a real-time PCR experiment in which the expression of a gene of interest in one sample (i.e., treated) is compared to expression of the same gene in another sample (i.e., untreated). The results are expressed as fold change (increase or decrease) in expression of the treated sample in relation to the untreated sample. A normalizer gene (such *GAPDH*) is used as a control for experimental variability in this type of quantification. The StepOne™ Real-Time PCR System (Applied Biosystems) was used to perform PCR amplification for mtDNA D-loop (C-tract) region. *GAPDH* was used as a ‘housekeeping gene’ to normalize all the threshold cycle (Ct) values. For each 10 µl reaction, 1 µl of unknown DNA was amplified containing 0.5 µl of each primer (20 pmol/µl), 5 µl of 2X SYBR Green Master Mix (Applied Biosystems) and 3 µl of nuclease-free water. The real-time PCR conditions consisted of initial denaturation and Taq polymerase activation at 95 °C for 10 min followed by 40 cycles of 95 °C for 45 s, 54 °C for 45 s and 72 °C for 1 min and followed by a melting curve analysis. Each measurement was repeated in triplicate, and a non-template control was included in each experiment. To determine the quantities of mtDNA and nuclear DNA (nDNA) present in the samples, the average threshold cycle (Ct) number values of the nDNA and mtDNA were obtained from each case. The level of mtDNA was calculated using the delta Ct (ΔCt) of average Ct of mtDNA and nDNA ($\Delta Ct = Ct_{mtDNA} - Ct_{nDNA}$) in the same well as an exponent of 2 ($2^{-\Delta Ct}$) (Mondal et al., 2013).

3.2.3.12 SNaPShot Multiplex System

The SNaPshot® Multiplex System is a primer extension-based method that enables multiplexing up to 10 SNPs (single nucleotide polymorphisms). It is used to screen and confirm SNPs, single locus fragment analysis, assess DNA methylation, Low-to medium throughput linkage and association studies, BACs fingerprinting, and screen for prion gene mutation. The SNaPshot system follows a straightforward protocol and uses infrastructure already existent in most clinical laboratories. This method consists of a multiplexed PCR step, followed by a single-base extension sequencing reaction, in which allele-specific probes interrogate loci of interest and are fluorescently labelled using dideoxynucleotides. These probes are designed to have different sizes and are subsequently resolved by electrophoresis and analyzed by an DNA sequencer. Thus, the identity of each locus is given by the position of its corresponding fluorescent peak in the spectrum, which is dictated by the length of the extension primer. The identity of the nucleotide(s) present at each locus is given by two parameters: the molecular weight and the colour of the fluorescently labelled ddNTPs added to the allele specific probes during the extension step (Figure 3.1). Thus, mutant and wild-type alleles can be distinguished based on the slightly different positions and on the distinct colours of their corresponding peak (Dias-Santagata et al., 2010).

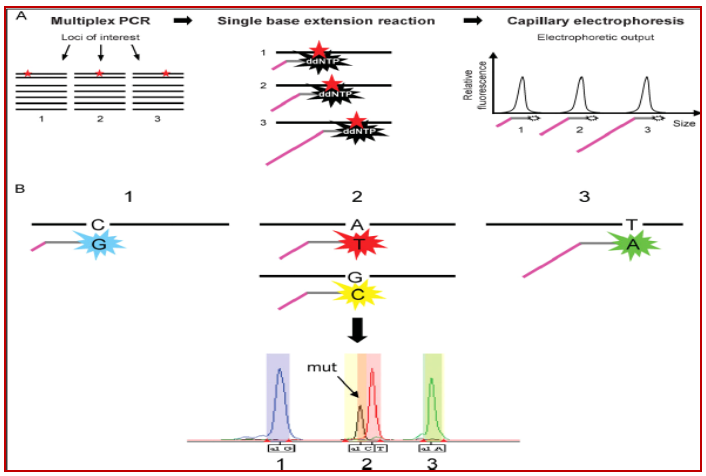


**Figure 3.1** Schematic representation of SNaPshot genotyping

3.2.3.13 Genotyping of *MECOM, TNFRSF19* and *CDK2BAS1* genes using SNaPshot assay

SNaPshot multiplex assay was used to validate the genotypes of *MECOM, TNFRSF19* and *CDK2BAS1* genes as predicted by conventional sequencing method. The SNaPshot reaction was carried out by using two sets of primers. First set is the external primer pairs (Table 3.1) that were used to generate the fragment containing the desired hotspot loci. Multiplex PCR was performed to get the desired products. Then the purified PCR products were used for SNaPshot single base extension reaction for genotyping the gene under study by another set of internal primers (Table 3.2).

**Table 3.2** External primer pair for SNaPshot

| SNP ID | Gene | Forward | Reverse |
|---|---|---|---|
| rs9510787 | *TNFRSF19* | CACTGTGTGAGAGAAAAGAAAGCC | TTCCCTTCCACCACTCACTTTT |
| rs6774494 | *MDS1-EVI1* | TTTTTGTCTGTATGCACCCAAAATCT | GAACTGACAAGCAATCATAATTCTTTCCAT |
| rs1412829 | *CDKN2BAS1* | CAGGTTGGTGACATCACCTGTT | TGCCCTTTTGCCTCATATTTATCTGAT |

**Table 3.3** Internal primers for SNaPshot

| SNP ID | Gene | Primer sequence |
|---|---|---|
| rs9510787 | *TNFRSF19* | AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGCTTCATAGTCTTAGAAGACAGC |
| rs6774494 | *MDS1-EVI1* | AAAATTTAATTTTTACAGGCCATGGCTT |
| rs1412829 | *CDKN2BAS1* | AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACTGTATTTTCTTTTGCCATTCCTCCA |

SNaPshot single base extension assay was performed using SNaPshot master mix with the above three sets of pooled primers (a final concentration of 5 pmol of the each primer was maintained) to amplify *MDS1-EVI1* (28 bp), *TNFRSF19* (58 bp), and *CDKN2BAS1* (74 bp) genes. Thermal cycler conditions were: initial denaturation step

at 95°C for 5 min, then 35 cycles at 95°C for 30 s, 50°C °C for 90 s and 72°C for 30 s, followed by final extension at 68°C for 10 min. Each 10 µl SNaPshot product was treated with 1 µl of Calf Intestinal Alkaline Phosphatase (CIAP) (1U/µl) at 37°C for 1 hr and further deactivation of CIAP was performed at 75°C for 15 minutes. An aliquot of each sample (1 µL) was combined with 0.5 µL of GeneScan™ 120 LIZ™ Size Standard and 9 µL Applied Biosystems® Hi-Di™ formamide. Samples were denatured for five minutes at 95°C and immediately chilled on ice (-20°C) before analysis on the Applied Biosystems® 3500xL Genetic Analyzer using POP-7™ polymer and the SNaPshot50_POP7xl run module.

**3.3 Bioinformatics analysis**

Different softwares such as ABI Sequence Scanner v1.0, ClustalX2, MEGA6, Bioedit and online servers such as NCBI Blast tool, Sequence manipulation suit, etc are used for the analysis of the sequences. The quality of the generated sequences (i.e. QV) were checked by Sequence Scanner v1 and SeqScape v2.7 (Applied Biosystems, Inc. USA) and further analyzed by nucleotide BLAST (Altschul, Gish et al. 1990) tool at the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/blast) to check percentage similarity of the developed sequences with the database sequences. The resultant sequences were then annotated and multiple sequence alignment was performed to detect specific SNP using BioEdit software (Hall 1999) and MEGA6 software (Tamura et al., 2013)

**3.4 Statistical analysis**

Various statistical tools like Student t-test, Fisher exact test and Pearson chi-square were used to compare various parameters. Both parametric logistic regression (LR) approach and non-parametric multifactor dimensionality reduction (MDR)

analysis were used to analyse high degree genetic and environmental interactions. False positive report probability (FPRP) was use to detect the consistency of LR and MDR analysis. To analyse the association of mitochondrial DNA copy number and NPC risk; mtDNA copies were categorized into quartiles based on the distribution among controls. A test for trend was calculated using the mtDNA copy number as a continuous variable. Spearman' rho test was used to correlate the nuclear DNA content and mitochondrial DNA content in peripheral blood in the study groups. Non-parametric Mann‑Whitney test was used to test if mtDNA content alteration is different in NPC cases, first-degree relatives (FDRs) and controls.

### 3.4.1 Logistic Regression Analysis

The association between the lifestyle habits and genetic factors and NPC risk was analyzed by calculating odds ratios (ORs), 95% confidence intervals (95% CI), and their corresponding *P*-values using Statistical Package for the Social Sciences (SPPS) software, version 16 for Windows. Further, gene-gene interaction and NPC risk, ORs were calculated for all the genotypes in combination. While analysis for gene–environment interaction were carried out stratifying the lifestyle habits. Departures from Hardy-Weinberg equilibrium were evaluated by comparing the expected frequencies to observed genotype frequencies using $\chi^2$ tests. P values less than 0.05 are considered statistically significant.

### 3.4.2 Multifactor Dimensionality Reduction (MDR) Analysis

The MDR software package (www. multifactordimensionalityreduction.org) was also used to detect the high level gene-gene and gene-environment interactions. MDR is a model free, non-parametric approach that can detect higher order interactions even in a small population by reducing the dimensionality of multi-locus information to

identify the polymorphisms or factors associated with an increased risk of disease. This helps in overcoming the limitations of low statistical power due to very high degrees of freedom when using logistic regression in studying higher order interactions. The best model for each order of interaction was selected by maximum cross validation consistency (CVC) and testing balanced accuracy (TBA). Interaction models showing highest TBA and CVC was further tested by 1000 folds permutation tests and $\chi 2$ test at 0.05% significance levels.

3.4.3 Interaction Entropy Graphs

The entropy-based analysis included in the MDR software package was used to determine synergistic and non-synergistic interactions among the variables. The graphs comprise of nodes containing entropy removed by individual variables and connections joining them pairwise showing entropy of interaction between them. Positive entropy signifies synergy and negative entropy indicate

3.4.4 False Positive Report Probability (FPRP)

Results generated by multiple comparisons analysis are often affected by the risk of being false positives. A major reason for this unfortunate situation is the strategy of declaring statistical significance based on a P values alone, particularly, any P values below 0.05. In order to detect the FPRP and the consistency of gene-environmental interactions (LR and MDR results), we used odds ratios and 95% confidence intervals from our analysis, observed p-values and power to detect ORs of 1.5 and 2.0 in a Bayesian approach (Wacholder et al., 2004). Considering a small sample size of our epidemiological data, the FPRP was computed using prior probabilities ranging from 0.25 to 0.001. Further, we considered findings with FPRP below 0.5 to be noteworthy result.