Materials and Methods
2.1. Study Population
   2.1.1. Screening and enrollment of schizophrenia patients and healthy controls
   2.1.2. Therapeutic interventions and clinical assessment

2.2. Candidate Gene and Variant Selection

2.3. Genotyping of genetic variants
   2.3.1. Genomic DNA isolation
   2.3.2. Homogeneity testing

2.4. Genotyping of prioritized genetic markers
   2.4.1. Illumina Golden Gate
   2.4.2. Sequenom

2.5. Genotyping quality control measures

2.6. Genotype-phenotype association analysis
   2.6.1. Population stratification
   2.6.2. Case control association study
   2.6.3. Family-based Test for Association Using TDT
   2.6.4. Combined Analysis of Case-control and TDT Studies
   2.6.5. Statistical Analysis
   2.6.6. Imputation

2.7. DNA binding assay of CTCF protein
   2.7.1. Gene cloning
   2.7.2. Polymerase chain reaction (PCR)
   2.7.3. Ligation of DNA termini
   2.7.4. Recombinant protein purification
   2.7.5. Gel shift assay

2.8. In silico Modeling
   2.8.1. HADDOCK Docking
   2.8.2. Molecular Dynamics Simulation
2.1. **Study Population**

2.1.1. **Screening and enrollment of schizophrenia patients and healthy controls**

A total of 736 individuals of North Indian ethnicity (351 cases and 385 controls) and 980 individuals of South Indian origin (436 cases, 401 controls, and 143 familial samples with 53 proband) were recruited from Clinical Services of Outpatients Department of Psychiatry Services, All India Institute of Medical Sciences (AIIMS), New Delhi and National Institute of Medical Health and Allied Sciences (NIMHANS), Bengaluru, respectively. Patients were diagnosed using the criteria described in the DSM-IV-TR (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition). The diagnoses were assigned based on a Structured Clinical Interview for DSM-IV-TR by two independent experienced psychiatrists. Other available resources, such as clinical course, medical records, and family history information were also documented. Further diagnosis was established on the basis of structured interviews by experienced psychiatrists using Schedules for Clinical Assessment in Neuropsychiatry (SCAN; WHO)(Wing et al., 1990) and OPCRIT 3.1 (MRC, Social, Genetic and Developmental Psychiatry Centre, Camberwell, South London) (McGuffin et al., 1994). These patients had received antipsychotic treatment at the same clinic. Patients were excluded if they had i) past or family history of any other mental illness or neurological disorder; endocrine disorders, and autoimmune diseases ii) substance dependence; iii) history of head injury; or iv) pregnancy at the time of enrollment. North Indian samples belong to Indo-European ancestry, whereas South Indian samples were of Dravidian origin Figure 2.1. Healthy individuals enrolled for the study were further interviewed to exclude any family history of psychiatric illness or antipsychotic drug exposure in their lifetime Voluntary informed consent were obtained after complete description of study to the participants were obtained from the patients and healthy individuals enrolled for the study, after explanation of the study protocol and their blood samples were collected for DNA isolation. Informed consent was obtained from respective family members if patient was not judged to be take his or her own decisions or think appropriately. This assessment of clinical judgment was performed by primary care physicians and senior psychiatric clinicians. This study was reviewed and approved by the Review Boards of the Hospitals and the Institute Ethics Committee.
Additionally, we conducted a pharmacogenetic based observational follow-up study in a natural setting. Patients were not controlled for their medications so drugs prescribed as usual in clinical practice and clinicians are blind to genotypes of the patients. Detailed clinical and treatment history including age of onset (AOO), duration of illness (DOI), medication, relapse, non-compliance and change of treatment were also documented through examination of hospital records and interviews of family members of the patients. Complete patient history was maintained for independent review. The age at onset of schizophrenia was defined as the age at first appearance of psychotic symptoms.

Figure 2.1: Representation of the linguistic map of India. Blue, brown, yellow, and pink backgrounds indicate regions where languages of predominantly IE, DR, AA, and TB lineages are spoken, respectively. In current study samples were derived from two linguistic regions i.e. IE and DR. (Indian Genome Variation, 2008)

2.1.2. Therapeutic interventions and clinical assessment

Schizophrenia patients were subjected to baseline clinical evaluations at the time of enrollment using the Clinical Global Impressions (CGI) scale (Busner and Targum, 2007). CGI is an assessment tool comprising of 3-item clinician-rated scale that measures severity of illness (CGI-S), global improvement or change (CGI-I) and
efficacy index (Appendix A). The CGI-S is rated on a 7-point scale, using a range of responses from 1 (normal) through to 7 (amongst the most severely ill patients). CGI-I scores range from 1 (very much improved) through to 7 (very much worse). Efficacy index ratings take account of both therapeutic efficacy and treatment-related adverse events and range between 0 (marked improvement and no side-effects) and 16 (unchanged or worse and side-effects outweigh the therapeutic effects). Each component of the CGI was rated separately to assess the patient’s status at the time of enrolment. This was followed by follow-up period of three months during which the patients were treated with atypical/typical antipsychotic drugs. The administered drugs included risperidone (2-10 mg/day), olanzapine (5-20 mg/day), clozapine (25-400 mg/day), ziprasidone (40-160 mg/day), quetiapine (1200 mg/day), aripiprazole (10-45 mg/day), amisulpride (100-600 mg/day), chlorpromazine (200-400 mg/day) and combination of two or more atypical/typical antipsychotic drugs. Since, in present study, clinicians were not controlling for any specific antipsychotic drugs, thus medications were prescribed according to chlorpromazine equivalence Table 2.1. It is required for switching between the antipsychotics and comparison of different antipsychotics for their efficacy, adverse events and adherence. In other words, to negate the fact that obtained associations do not because of high or inadequate dose of any particular drug treatment type (Am J Psychiatry, 1997; Chue et al., 2005; Nayak et al., 1987; Schooler and Levine, 1976; Woods, 2003).

Table 2.1: Chlorpromazine (100mg/day) equivalent doses for antipsychotics prescribed in present study.

<table>
<thead>
<tr>
<th>Generic Name of antipsychotics</th>
<th>Equivalent dose (mg/day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risperidone</td>
<td>2</td>
</tr>
<tr>
<td>Olanzapine</td>
<td>5</td>
</tr>
<tr>
<td>Clozapine</td>
<td>100</td>
</tr>
<tr>
<td>Ziprasidone</td>
<td>60</td>
</tr>
<tr>
<td>Quetiapine</td>
<td>75</td>
</tr>
<tr>
<td>Aripiprazole</td>
<td>7.5</td>
</tr>
<tr>
<td>Amisulpride</td>
<td>100</td>
</tr>
</tbody>
</table>
Materials and Methods

Complete patient history, their medication, and follow-up of all the patients (to ensure their stability, relapses and change in treatment) were maintained for independent review. Following a three-month antipsychotic treatment period, the patients were again assessed for treatment response using CGI.

A Total of 419 patients from South India and 323 patients of North India were completed three months follow-up, (Table 1b). It is expected that severely ill patients have poorer treatment outcome, especially if they don't respond to several months of treatment or if they are not responding, their symptoms may become more severe. Further, patients who scored 2 or less in CGI-Improvement (CGI-I) or drop of 2 from baseline score were described as complete responders (CR) towards antipsychotic treatment. Patients with CGI-I score 3 or more were grouped as incomplete responders (IR) (Gupta et al., 2009; Gupta et al., 2012; Gupta et al., 2013; Kaur et al., 2014a; Kaur et al., 2014b). Existing literature supports evaluation period of three months is sufficient to understand antipsychotic response (Lehman et al., 2004) (Meltzer et al., 2008). Thus for present study, patients were followed for three months during which they were treated with antipsychotic drugs (Fijal et al., 2009; Tybura et al., 2012). According to baseline severity score, patients were broadly categorized into low severity (LSG; CGI-S score ≤3) and high severity groups (HSG; CGI-S ≥4). (Leucht and Engel, 2006) reported CGI may have similar sensitivity as the BPRS to evaluate various antipsychotic drugs efficacy, hence for present study we considered Clinical Global Impressions Severity scores (CGI-S) to evaluate response of patients towards drugs. Clinicians have critically evaluated and analyzed the every aspect of the individual before rating them on CGI scale for their response towards particular drug. If the severity of illness symptoms continued (CGI-S > 3), treatment was modified according to guidelines and they were evaluated again after 12 weeks. A fall in CGI-S scores by two points was qualified as complete responder. Those who continued to have scores of 3 or above were classified as incomplete responders to antipsychotic treatment.

Among South India patient 192 (45.60%) patients were classified into LSG, among them 142 (33.57%) were CR and 50 (12.06%) were IR. And 227 (54.37%) patients fell into HSG group with 67 (16.07%) as CR and 160 (38.30%) as IR. From North India group 211 (45.60%) patients were classified into LSG, among them 146
(33.57%) were CR and 65 (12.06%) were IR. And 112 (34.68%) patients fell into HSG group with 11 (16.07%) as CR and 102 (38.30%) as IR. The protocol for antipsychotic drug response evaluation has been represented in Figure 2.2.

Figure 2.2: Study design for evaluating antipsychotic drug response.

2.2. Candidate Gene and Variant Selection

The advantage of performing a study focused candidate genes is cost efficiency. There is requirement of relatively less number of markers to capture most of the common variation in the candidate region. Candidate gene/region studies should follow genomic linkage or association studies that have already implicated the region. Linkage studies indicate evidence of rare variants with big effects, whereas association studies are powered to pick up common variants with modest effect size. However, in candidate gene study where selection is made entirely on the basis their biological significance, without any previous statistical evidence that the region they are located in is implicated, have had very low success rates even when well-designed (Hirschhorn et al., 2002). Alternatively, candidate genes can be selected from biological pathways which harbour other previously associated risk loci. The selection
Materials and Methods

of SNPs in or around the candidate gene region should include functional elements present in region. As genetic variant may affect the gene expression or can alter the protein functioning. Regulatory regions such as promoters and enhancers are essential for controlling the extent to which the gene is transcribed. Polymorphisms in coding region and regulatory regions are less frequent than variations in intron. This is explained by the fact that slightest variation in fine-tuned biological system will likely have a detrimental effect on the proper functioning of process.

In the current study, the main criterion for selection of candidate gene/pathway were based on their involvement in disease and therapeutics NRG1-ERBB signalling pathway, neuroactive ligand receptor interaction, glutamate signalling, and their downstream signalling kinase gene, and for follow-up from previous GWAS conducted in patients with schizophrenia and antipsychotic response (Consortium, 2011; Need et al., 2009; Shi et al., 2009).

Most of the previous studies take Serotonin and Dopamine receptors into consideration. As these receptors has important role in regulating brain physiology by interacting with neurotransmitter directly. Recent evidences suggested potentiality of neuregulin signaling pathway in development of new generation antipsychotic (Deng et al., 2013). Confirmations that second generation antipsychotic antipsychotics have neurogenic action in brain region encourages to study the involvement of neurodevelopmental genes in antipsychotic response in our population (Newton and Duman, 2007). Therefore for the present study, an integrative approach using both functional and positional strategies as well as extensive literature mining was adopted for prioritizing candidate genes which might influence response to antipsychotic drugs. The function-dependent strategy is critical in identifying genes which might have a functional role in modulating the biological functional environment of the drug–target interactions (Roden and George, 2002). The biological significance of these genes and importance of functional genetic polymorphisms in these genes has already been explained in details in the review of literature section.

We utilized custom set of 1536 SNPs, of which 984 were prioritized from 40 candidate genes based on gene coverage, functional prediction, and prior reports. Also, 552 neutral markers which are not located in region linked to psychiatric illness and in genes to be implicated in ay brain disease, neuronal development and neuronal
functioning were selected for homogeneity test. These markers were selected from 4991 neutral markers form Affymetrix 50K dataset (Jha et al., 2012). SNPs were prioritized from these candidate genes using literature reports. Further, additional SNPs were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/) and UCSC Genome Browsers (http://genome.ucsc.edu/) for uniform gene coverage. All the SNPs were subjected to in silico functional analysis using SIFT, PolyPhen-2, Haploreg and RegulomeDB (online ENCODE data repositories) to assess their putative functional role on gene expression and activity, if any. NCBI Genome Browser has been used to retrieve the gene and SNP details.

2.3. Genotyping of genetic variants

2.3.1. Genomic DNA isolation

Genomic DNA was isolated from the peripheral blood leukocytes using a modification of a salting out procedure (Miller et al., 1988). Approximately, 8-10 ml blood was drawn from patients as well as healthy controls by venepuncture and collected in vacutainers containing acid citrate dextrose (ACD) buffer. Red blood cell lysis buffer (RLB, 1X) was added to the blood in a falcon to make the final volume 50ml (Appendix I). The suspension was mixed by inverting the falcons several times until it became translucent and was further incubated for 20 minutes at room temperature. This was followed by centrifugation at 2500 rpm for 15 minutes at room temperature and the supernatant was discarded in sodium hypochlorite solution. Further, 15ml RLB was added to the pellet. The suspension was mixed by inverting the tubes several times, followed by centrifugation at 1000 rpm for 10 minutes. The supernatant was discarded in sodium hypochlorite and the pellet was re-suspended in 12ml of nuclei lysis buffer (NLB), followed by addition of 0.8ml of 10% sodium dodecyl sulphate (SDS) and 50µl Proteinase K (20µg/µl). After incubation at 65°C for 2 hours, the digested protein product was precipitated by the addition of 4ml of 6M sodium chloride (NaCl) solution. After centrifugation for 30 minutes at 3500 rpm at room temperature, the supernatant was transferred to a falcon. Absolute ethanol was now slowly added to the supernatant to spool the DNA. The spooled DNA was further washed with 70% ethanol twice, air-dried, and dissolved in Tris-Ethylendiaminetetraacetic acid (TE) buffer. The quantity of the DNA was estimated
using a spectrophotometer to determine the optical density at 260nm and 280nm. DNA quality was assessed using the 260nm/280nm ratio. The stock solution of the DNA was stored at -20ºC. The composition of buffers and reagents are provided in Appendix B.

2.3.2. Homogeneity testing
We selected 552 neutral markers to test presence of substructures in the studied population. These markers are not located in regions linked to psychiatric illness or known to implicate in brain diseases, brain development and functioning were selected for present study. These markers were selected from 4991 neutral markers from Affymetrix 50K dataset (Jha et al., 2012).

2.4. Genotyping of prioritized genetic markers
Genotyping was performed at Genotyping and Sequencing Facility of CSIR-IGIB using customized array of Illumina Genomestation using GoldenGate technology according to manufacturer’s guidelines using 5µl of DNA with concentration of 5ng/µl. For genotyping quality control measures, 88 samples of DNA were provided as duplicates for clustering and reproducibility. Primary genotyping data analyses were executed with GenomeStudio software 2010.3, and followed by visual inspection and assessment of data quality and clustering. To rule out any genotyping error, 15% of the total samples were re-genotyped randomly using MALDI-TOF mass spectrometry platform (SequenomTM, Inc., SanDiego,CA).

2.4.1. Illumina Golden Gate
The GoldenGate Assay allows for a high degree of loci multiplexing (1536-plex) during the extension and amplification steps.

Assay Design: The Assay Design tool (ADT) provides a convenient method to design a custom panel of 1536 SNPs on Illumina GoldenGate Assay panel. The entire process of selecting and submitting a list of selected variant and loci to Illumina is via i.com.illumina.com, using user specific ID. With support of illumina scientist ADT file is processed, in response the score output file is generated. This score output file
Materials and Methods

contain detailed information for each marker. The Final score section in output indicates the success for designing a given assay.

Assay overview shown in Figure 2.3: The DNA sample used in this assay is activated for binding to paramagnetic particles step-1. This activation step is a robust process that requires a input of DNA (250ng at 50ng/μl). Assay oligonucleotides, hybridization buffer, and paramagnetic particles are then combined with the activated DNA in the hybridization step-2. Three oligonucleotides are designed for each SNP locus. Two oligos are specific to each allele of the SNP site, called the Allele-Specific Oligos (ASOs). A third oligo that hybridizes several bases downstream from the SNP site is the Locus-Specific Oligo (LSO). All three oligonucleotide sequences contain regions of genomic complementarity and universal PCR primer sites; the LSO also contains a unique address sequence that targets a particular bead type. During the primer hybridization process, the assay oligonucleotides hybridize to the genomic DNA sample bound to paramagnetic particles. Because hybridization occurs prior to any amplification steps, no amplification bias can be introduced into the assay. Following hybridization, several wash steps are performed, reducing noise by removing excess and mis-hybridized oligonucleotides. Extension of the appropriate ASO and ligation of the extended product to the LSO joins information about the genotype present at the SNP site to the address sequence on the LSO step-3. These joined, full-length products provide a template for PCR using universal PCR primers P1, P2, and P3 step-4. Universal PCR primers P1 and P2 are Cy3 and Cy5-labeled. After downstream-processing step-5 the single-stranded, dye-labeled DNAs are hybridized to their complement bead type through their unique address sequences step-6. Hybridization of the GoldenGate Assay products onto the Array Matrix or BeadChip allows for the separation of the assay products in solution, onto a solid surface for individual SNP genotype readout step-7. After hybridization, the BeadArray Reader is used to analyze fluorescence signal on the Sentrix Array Matrix step-8. Analyzed using genome-studiosoftware for automated genotype clustering and calling step-9.
2.4.2. Sequenom

The sequenom iPLEX assays (Increased Plexing Efficiency and Flexibility for MassARRAY System) enabled high throughput genotyping and analysis of SNPs through the application of matrix assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) Figure 2.4. The iPLEX method involves multiplex PCRs followed by single base primer extension reactions, resulting in allele-specific differences in mass between extension products which are analyzed by MALDI TOF MS. The method comprises of three basic steps described below:

Assay overview
(i) Assay Design: The SNP sequences (150-200 bp sequences flanking the SNP) were obtained from Ensembl Genome Browser (www.ensembl.org) and
MassARRAY Designer software (Sequenom Inc., San Diego, USA) was used for designing both the PCR and iPLEX single base extension primers for each SNP based on the sequences.

(ii) Genotyping protocol: PCR amplification reactions were performed in 5 μl reaction volume using 0.625μl PCR buffer (10X), 0.325μl MgCl₂ (25mM), 0.1μl dNTPs (25mM), 1μl primer mix (500nM), 0.1μl Hotstar Taq polymerase enzyme (5U/μl), 1.85μl nanopure water and 1μl genomic DNA (5-10ng/μl). The PCR thermocycling conditions involved an initial denaturation step of 5min at 95°C, followed by 45 cycles of 20sec denaturation at 95°C, 30sec annealing at primer-pair specific temperature, extension at 72°C for 1min; and a final extension at 72°C for 3min. As PCR amplification was carried out for multiplex reactions, PCR reagent concentrations and conditions were optimized for each plex. The amplicons were then subjected to SAP treatment to dephosphorylate unincorporated dNTPs, where 2μl SAP mix (0.17μl SAP buffer (10X), 0.3μl SAP enzyme (1U/μl), 1.53μl nanopure water) was mixed with 5μl amplicons and incubated at 37°C for 20 minutes, followed by 5 minute incubation at 85°C. The purified amplicons were then used to prepare the iPlex cocktail mix (0.2μl 10X iPLEX buffer, 0.2μl iPLEX termination mix, 0.804μl primer mix, 0.041μl iPLEX enzyme and 0.755μl nanopure water) for iPLEX primer extension according to the manufacturer’s protocol. (Sequenom Inc., San Diego, USA). iPLEX reactions were carried out using a 200-short-cycle program uses two cycling loops of 40 and 5 cycles as described below:

```
94°C for 30 seconds
94°C for 5 seconds
52°C for 5 seconds
5cycles
80°C for 5 seconds
52°C for 5 seconds
5cycles
72°C for 3 minutes
80°C for 5 seconds
5cycles
4°C forever
```

The iPLEX reaction products were then desalted using clean resin (Sequenom Inc., San Diego, USA) and dispensed onto a SpectroCHIP bioarray using a nanodispenser. The SpectroCHIP arrays are placed into the MALDI-TOF mass spectrometer and the mass correlating genotype is determined in real time. A SpectroCHIP is typically
Materials and Methods

processed in 45-60 minutes. MassARRAY Workstation software was used to process and analyze iPLEX SpectroCHIP bioarrays. These steps were carried out by trained individuals.

(iii) MALDI-TOF MS analysis: MassARRAY Typer software version 3.4 was utilized to visualize the SNP allele peaks (mass spectra data) and assess the data quality, assay success rate and genotype calls for each SNP. Every SNP was manually checked for quality by looking at the spectrum and cluster plots. The calling algorithm, Caller version 3.4 (Sequenom Inc., San Diego, USA) was used to create genotyping calls. The genotype calls are categorized as no calls, low mass homozygous, heterozygous, or high mass homozygous. No calls made by software in case of low probabilities of either allele were manually assigned after thorough inspection of all the SNP spectra. For SNPs with low genotype calls or improper cluster plots, genotyping was repeated using SNaPshot and sequencing method.

![Image](image.png)

Figure 2.4: Overview of MALDI-TOF assay (adapted from Sequenom.Inc)

2.5. Genotyping quality control measures

The QC of markers in candidate-gene studies is more comparable to that for the GWA study approach, as similar numbers of cases and controls are involved. It is extremely important to examine the failure rate of markers included in the candidate-gene study
and exclude those with a high failure rate. Detection of deviations from HWE in controls is still a relevant method for checking genotyping quality.

Limitation in study design and errors in genotype calling can introduce systematic biases into genetic case-control association studies, leading to an increase in the number of false-positive and false-negative associations (Anderson et al., 2010). Many such errors can be avoided through a careful selection of case and control groups and attentive experimental practices. However, even having appropriate case-control association study design lack in a thorough assessment of data quality-including homogeneity testing or whether both case and control are from same population should be undertaken during sample ascertainment. Such assessments allow the identification of inferior markers and samples, which should be removed before taken them into further analysis to reduce the number of false-positives and false-negatives.

To ensure consistent and high-quality genotyping data, more than 15% of the total samples were re-genotyped using Sequenom platform. The duplicate genotypes were analyzed and the genotyping error rate was checked. During the initial quality check (QC), loci were screened and excluded from analysis for following reasons: poorly defined clusters, GenTrain score < 0.6, more than 60% genotyping failure, significant differences in missingness between cases and controls, excessive Mendelian and replication errors and samples having all the heterozygous calls (Anderson et al., 2010). In addition, failed genotyping reactions were repeated. Further, SNPs or samples with an overall call rate below 90% were excluded, as were SNPs with unusually low spectral peak intensities. Hardy-Weinberg Equilibrium (HWE) was assessed for all the SNPs that passed initial QC. SNPs deviating from HWE were check by a goodness of fit test (df=1) with threshold of significance p<0.0001 in controls. SNPs with minor allele frequency (MAF) < 1% and non-polymorphic nature were not included in further analysis. Genotype concordance was ~98.8% for Illumina and ~94.6% for Sequenom platform.
2.6. Genotype-phenotype association analysis

2.6.1. Population stratification

One of the largest sources of confounding in association studies using unrelated individuals holds from population stratification, which occurs when the study population contains multiple subgroups of individuals with distinct genetic backgrounds (usually the result of including multiple racial or ethnic subgroups into a single study population). This becomes problematic and leads to systematic false positive and false negative errors (Marchini et al., 2004) when two conditions apply: (1) The subgroups differ with respect to the frequency of an allele, and (2) The subgroups have different average values for the quantitative outcome of interest (or differ in the frequency of the occurrence of an event if the outcome under investigation is categorical). If a truly irrelevant polymorphism was more common in the subgroup that, for some other genetic or non-genetic reason, has a higher or lower average value for the trait being studied, then the allele will appear associated with trait in this dataset. The association here would be purely artifactual, resulting from the failure to adjust for population stratification, in addition to causing excessive type I error inflation, population stratification could also obscure true genetic associations (Marchini et al., 2004).

Statistical methods for detecting and controlling for population stratification- To rule out the possibility of population stratification, patients and unrelated healthy individuals were recruited from the same ethnicity, it rarely eliminate possibility of stratification. Another way to resolve this problem is to use family-based designs, which are robust to population stratification (Cardon and Bell, 2001; Cardon and Palmer, 2003). Rather than associating frequency of alleles across families, family based designs typically link or associate outcomes to regions of genetic variation by following alleles through meioses within families, which are robust to confounding by population substructure. To deal with these challenges, statistical methodology has been developed and implemented into software to aid in detecting and adjusting for population stratification in genetic association studies. The inflation factor can also be used to adjust for stratification. An inflation factor close to one would indicate stratification is not present in the data (Devlin and Roeder, 1999; Reich and Goldstein, 2001). Structured association (Pritchard et al., 2000), implemented in the
**Materials and Methods**

STRUCTURE software (Pritchard and Donnelly, 2001), uses genotype data to infer population structure, and then performs tests of association within each inferred subpopulation. Investigators may also use STRUCTURE to identify individual samples that do not cluster with the majority of the samples. These samples may then be eliminated from the analysis. Because the risk of confounding by population stratification increases with sample size (Marchini et al., 2004), and because extremely large sample GWAS are becoming increasingly common, another method has been developed that utilizes large samples and thousands of markers throughout the genome to correct for population structure. Eigenstrat (Price et al., 2006) uses principal components analysis (PCA) to explicitly detect and correct for population stratification on a genome-wide scale in large sample sizes in a computationally efficient manner. A thorough discussion of stratification, principal component analysis, and the eigenvalues can be found (Patterson et al., 2006).

Objective of the Principal Component Analysis based technique was to first determine the different patterns of variation over the subjects from within the markers, after that those variation are resolve using each marker.

**Procedure:**

1. The average value for the marker across all samples is determined.
2. The average value is subtracted from all values for the marker to recenter the entire data to zero resulting in pattern of variation for the particular marker.
3. Pattern is normalized by calculating theoretical standard deviation of the marker’s data at Hardy-Weinberg Equilibrium (HWE). This normalized marker pattern over the entire set of subjects (individuals) are summed into "Wishart" matrix \( (m \times n) \) or \( X^T X \).

The eigenvectors and eigenvalues of Wishart matrix are its components, and eigenvectors corresponding with eigenvalues are called its principal components. Generally, first few principal components correspond directly to the stratification patterns. Therefore, the final two steps of stratification correction through the “EIGENSTRAT” PCA technique are to find the top \( k \) principal components (where you select \( k \)) and remove these component/patterns from both the marker data and the dependent variable using a vector-analysis related technique.
A recent report using large-scale simulation studies to compare methods for correcting for population stratification examined all of the above-mentioned techniques, and found that PCA-based methods (as implemented in Eigensoft) outperformed both genomic control and structured association in terms of maximizing power, controlling false positives maintaining in computational efficiency (Zhang et al., 2008).

Population structure was investigated by Principal component analysis performed for set of neutral markers with Eigenstrat (Price et al., 2006). Using admixture model, allele frequency corelated. Principal component analysis was performed in north and south Indian samples respectively to assess population structure and identify, outliers with 552 independent SNPs using SMARTPCA as implemented, in EIGENSOFT version 3.0. The SMARTPCA parameters used for removing the outlier included \( \sigma \) (sigma) threshold of 6 with 5iterations, along first 10 principal components. Genotype frequencies of each of the neutral markers were compared between the patient and healthy control groups, and between complete and incomplete responders using the Pearson \( \chi^2 \) tests for association to test for population stratification. The \( p \)-value of the observed test statistic was estimated as the fraction of 10,000 simulated test statistics that exceeded the observed value using the STRAT program (Pritchard and Rosenberg, 1999). The sum of the test statistics for each locus was then computed with the number of degrees of freedom (df) being equal to the sum of the number of df of the individual loci as implemented in STRAT program. Neutral markers were analyzed using the STRUCTURE software was used and admixture model was assumed (Pritchard et al., 2000). According to this model individual, \( i \) have inherited some fraction of his/her genome from ancestors in population \( K \) and it computes the proportion of the genome of an individual originating from each inferred population posterior mean estimates known as \( Q \). This model is recommended due to its flexibility for dealing with real population complexities. We assumed \( K=2 \) and model parameters includes length of burning period, number of Markov Chain Monte Carlo (MCMC) reps after burning, and iterations as 100, 100 and 100 respectively were considered.
2.6.2. Case control association study

A genetic association analysis is fundamentally similar to any other statistical association analysis. The main objective is to establish an association a disease trait (dichotomous or continuous variable) and a genetic marker (genotype, a categorical variable). The disease trait is often dichotomous, or a quantitative measure like obesity, or age of onset of a disease. The genetic marker can be a known or unknown to cause any effect on DNA, like coding or affecting regulatory factor binding.

Genetic association analysis is of two different study designs; one which uses only unrelated subjects and the other which uses families that have at least two members with genetic marker data. Family designs have several advantages and disadvantages (describe in the next section) and are an important class of studies to examine transmission of disease causing marker from parent to offspring.

In the case-control design we select diseased cases and healthy controls, from similar populations, in terms of age, sex, ancestry etc. Sometimes a particular mode of inheritance is known from the previous association studies or biological considerations. But this case is often rare for complex disorder like studied schizophrenia. Most often there are two alleles, and three genotypes to study relationship between disease and genotype under different inheritance models, the recessive, the dominant, and the additive models. The most popular assumption is an additive mode, which assumes that the risk associated with the heterozygote genotype is intermediate between the two possible homozygotes (N.M. Laird, C. Lange, 2011).

In the present study, case control association design in schizophrenia cases and non-diseased control from the North India and South India population was utilised as mentioned in beginning of the section. Genotype effect for disease state were assessed using general additive model regression methodology, adjusted for age, gender, and ancestry factors. Our samples pool consists of North Indian samples which belong to Indo-European ancestry, whereas South Indian samples were of Dravidian origin. These two populations are mutually exclusive and are from different part of country based on demographic and geography. Voluntary informed consent was obtained after complete description of study to the participants. In case patients were unable to give their consent their family members were asked to provide consent on behalf of the patient. This current study was reviewed and approved by the Review
Boards of the both the participating Hospitals and the Institute Ethics Committee of CSIR-IGIB.

2.6.3. Family-based Test for Association Using TDT
The TDT is a commonly used method for controlling for confounding due to population stratification in association studies. The method was originally proposed as a follow-up test for disease-marker association obtained in population-based case-control studies (Spielman and Ewens, 1996). It avoids the problem of population stratification by using un-transmitted alleles as internal controls for transmitted alleles from heterozygous parents to affected offspring.

2.6.4. Combined Analysis of Case-control and TDT Studies
The odds ratio is a commonly used measure of strength of association since its logarithm is asymptotically normally distributed. If the odds ratios of the association of a genetic marker with a disease are obtained in both case-control and TDT studies, then a weighted analysis can be performed to estimate the combined odds ratio and its associated standard error (SE). For case-control studies, estimates of the odds ratio and standard errors are obtained as described above. For TDT studies, the odds ratio of disease risk can be estimated using the proportion.

Having obtained estimates of the logarithm of the odds ratios and their associated standard errors in case-control and TDT study designs, an estimate of the combined odds ratio and its associated SE can be obtained using a weighted analysis method (Hedges & Olkin, 1985).

A useful method to confirm gene-associations detected in case-control studies is the Transmission/Disequilibrium Test (TDT). The TDT approach was introduced by (Spielman et al., 1993) and provides a simultaneous test of linkage and gene-association. This is particularly advantageous as the statistic is unaffected by population admixture. A convenient summary statistic of the magnitude of the disease-marker association can be estimated from either case-control studies or the TDT, enabling comparisons of results for the same disease-marker association (Mitchell, 2000). Furthermore, it would be useful to obtain a combined estimate of the disease-marker association if the diseases in the two studies are comparable and both
populations share the same genealogy. In this case a fixed-effects model would be applicable for a combined analysis. The combined estimate of a disease-marker association can be used to contrast results from the two studies and provides an overall picture of the effect size attributable to such a polymorphism. Let $\phi^1$ denote the logarithm of the odds ratio in case a case-control study with variance $\sigma^2_1$ and $\phi^2$ the logarithm of the odds ratio in the TDT with variance $\sigma^2$. Since the logarithm of the odds ratio is asymptotically normally distributed, estimates of the combined logarithms of the odds ratios can be obtained as:

$$\hat{\Psi} = \frac{\sum_{i=1}^{2} w_i \phi_i}{\sum_{i=1}^{2} w_i}$$

with variance given as:

$$\text{Var}(\hat{\Psi}) = \frac{1}{\sum_{i=1}^{2} w_i}$$

Where $w_i = 1/\sigma^2_i$ are the weights. Significance of the combined odds ratio estimate can be examined using the test statistic

$$X^2_{\Psi} = \frac{(\hat{\Psi} - \Psi_0)^2}{\text{Var}(\hat{\Psi})}.$$ 

The test statistic follows a chi-square distribution with 1 degree of freedom under the null hypothesis of no association (i.e. $\psi_0 = 0$). In addition, homogeneity of the odds ratios from the two studies is accessed via the test statistic

$$X^2_H = \sum_{i=1}^{2} w_i (\phi_i - \hat{\Psi})^2 \sim \chi_1^2.$$ 

A Transmission Disequilibrium Test (TDT) analysis was performed in South Indian trios using PLINK and overall estimates of allelic effect were calculated by case-control studies and TDT jointly analyzed to obtain a combined estimate of disease-marker association. The combined analysis recommended in published literature (Kazeem and Farrall, 2005).

In the South Indian data set, firstly we performed a family based association study, using trio design; analysis was performed using the Transmission Disequilibrium Test (TDT). Later, we performed combined analysis of unrelated case-control samples.
comprising 435 cases and 393 healthy individual and familial samples in trios setting comprising 143 individuals with 53 probands, enrolled in the study.

2.6.5. Statistical Analysis

In the present study, GenomeStudio Genotyping Module v1.0 was used to analyse SNPs of the Illumina custom panel for DNA sample validation. All data management and quality control assessments were performed with PLINK 1.07 (Purcell et al., 2007); http://pngu.mgh.harvard.edu/purcell/plink/) unless mention. Studies by Indian Genome Variation Consortium and pioneering work by Reich and colleagues had shown that Indian populations demonstrate the presence of north and south gradient (Reich et al., 2009). The two studied population are ancestrally different. Workers in the field had provided the evidence of two ancient and genetically divergent populations. Studies had documented a high level of population structure in Indian population. So, from the point of view of above studies it is not suggestive to pool the data from two populations. This may result in spurious association due to systematic ancestry differences in two populations. Hence, we have performed analysis separately for both the populations. Results from the two study data sets were compared by meta-analysis using PLINK Software. Heterogeneity was evaluated with the Cochran Q statistic test in the contributing data sets. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated using R script (appendix for the formulae). Linkage Disequilibrium (LD) was estimated among SNPs by $r^2$ value for the genotype data of all markers in healthy controls using Tagger algorithm in Haploview program version 4.1 (Barrett et al., 2005). Identification of tagSNPs was performed using Tagger program.(de Bakker et al., 2005) A Transmission Disequilibrium Test (TDT) analysis was performed in South Indian trios using PLINK software and overall estimates of allelic effect were calculated as mention above (Kazeem and Farrall, 2005). Chi square test was used for the association analyses of SNPs with drug response under additive model of inheritance. Odds ratio (OR) and confidence interval (95% CI) were calculated for each marker.
2.6.6. Imputation

Entire data was phased before using PHASE software, based on the Bayesian algorithm developed by Stephens and colleagues (version 2.1) for imputing the missing genotype in the samples based on 'in-samples' criteria. Parameter values of 100 iterations, a thinning interval of 10, and a burn-in value of 100 in the MCMC simulations were used (Stephens et al., 2001).

2.7. DNA binding assay of CTCF protein

The electrophoretic mobility shift assay (EMSA) is an established method to detect DNA binding proteins. The principle being that a nucleic acid with protein bound, has less mobility through a gel matrix than free nucleic acid. One of the primary applications of the EMSA is the detection of transcription factors and other sequence specific DNA binding proteins. In the current work DNA binding activity of CTCF-transcription factor was assessed using EMSA using DNA region of \textit{HTR3A} promoter region with single nucleotide alteration at 42 base pair upstream of the gene.

2.7.1. Gene Cloning

Total Human RNA was isolated from human hepatoma cell line (HepG2) and used for cDNA preparation by High capacity reverse transcription cDNA synthesis kit (Applied Biosystem) as per manufacturer's instructions. cDNA was PCR amplified using CTCF specific primers (FP: \texttt{CGCGGATCCCTTTGCAGCCACGGAGAG}, RP: \texttt{CCGCTCGAGAACACAGCCAGAAGTCC}) and Phusion High-Fidelity DNA Polymerase (ThermoScientific) that was further purified with QIAquick PCR purification kit (Qiagen). The purified PCR product and the vector (pET28-His\textsubscript{10}-Smt3) were double digested with BamHI and XhoI restriction enzymes. The digested DNA was gel purified using QIAquick Gel Extraction Kit (Qiagen) and ligated using T4 ligase enzyme (ThermoScientific) as per manufacturer's instructions. The ligated product was chemically transformed into \textit{E. coli} DH5\textalpha{} cells. The clones were identified by restriction digestion and sequencing.
2.7.2. Polymerase chain reaction (PCR)

PCR reactions were performed using previously described methods (Sambrook et al., 1982). Total Human RNA was isolated from human hepatoma cell line (HepG2) and used for cDNA preparation by High capacity reverse transcription cDNA synthesis kit (Applied Biosystem) as per manufacturer’s instructions. All PCR reactions were performed using Pfu DNA polymerase cDNA was PCR amplified using CTCF specific primers (FP: CGCGGATCCCTTGTGCAGCCACGGAGAG, RP: CCGCTCGAGAACACAGCCAGAGAGTCC) according to Table 2.2.

Table 2.2: A PCR amplification reaction composition

<table>
<thead>
<tr>
<th>PCR components</th>
<th>Working concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template cDNA</td>
<td>50 ng</td>
</tr>
<tr>
<td>Pfu DNA polymerase buffer</td>
<td>1 X</td>
</tr>
<tr>
<td>Oligonucleotide forward primer</td>
<td>5 pmoles</td>
</tr>
<tr>
<td>Oligonucleotide reverse primer</td>
<td>5 pmoles</td>
</tr>
<tr>
<td>dNTP mix</td>
<td>0.25 mM</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>1.5 mM</td>
</tr>
<tr>
<td>DMSO</td>
<td>6 %</td>
</tr>
<tr>
<td>Pfu DNA polymerase</td>
<td>2.5 U</td>
</tr>
<tr>
<td>Sterile double distilled H₂O</td>
<td>Upto 50μl</td>
</tr>
</tbody>
</table>

For the amplification of different genes, a typical amplification reaction comprised of:
1. Initial denaturation at 95°C for 5 min.
2. 30 Cycles of: denaturation at 95°C for 1 min, annealing at 55°C - 60°C for 1 min and extension at 72°C for 1 min/1000bp.
3. Final extension at 72°C for 10 min.

PCR products were resolved on 1% agarose gel and purified using gel extraction kit. The purified PCR product and the vector (pET28-His₁₀-Smt3) were double digested with BamHI and XhoI restriction enzymes. The digested DNA was gel purified using QIAquick Gel Extraction Kit (Qiagen) and used for ligation.
2.7.3. Ligation of DNA termini
Ligation reactions were carried out in a volume of 10 µl at 16°C for 16 hr. The reaction mixture contained ~100 ng of the digested pET28-His10-Smt3 expression vector, CTCF DNA fragment at 1:3 (vector:insert) molar concentrations, 1X T4 DNA ligase buffer and 10 U of T4 DNA ligase. The ligation reaction products were then used to transform competent cells of *E. coli* DH5α and transformants were selected on LB agar plates supplemented with kanamycin.

2.7.4. Recombinant protein purification
For over-expression of His6-tagged CTCF, pET28-His10-Smt3-CTCF was transformed in BL21 (DE3) cells and transformants were grown in LB medium containing kanamycin at 37°C until OD<sub>600</sub> = ~0.6. The cultures were induced with 1 mM IPTG and incubated for 12-14 hr at 16°C followed by cell harvesting. All the purification procedures were performed at 4°C.

For purification, the harvested cells were dissolved and sonicated in lysis buffer. After removing the cytosolic fraction by centrifugation, the protein from inclusion bodies was recovered using solubilization buffer. After centrifugation (16000 rpm, 30 min, 4°C), the supernatant was incubated with Ni<sup>2+</sup>-NTA affinity resin (Qiagen) pre-equilibrated with Buffer A1. The column was washed extensively with buffers A1, B1 and C1. Desired His6-tagged CTCF was obtained using elution buffer E1. Purified protein was run on 10% SDS-PAGE and analyzed by Coomassie staining.

2.7.5. Gel shift assay
For protein-DNA binding assay, three primers were used. Primer 1 (5’-GTGGGCCTCGCTTGAGCACT-3’), primer 2 (5’- GTGGGCCTGC[T]CCTGAGCACT-3’) and primer 3 (5’- GTGGGCCTCGC[CH₃]CCTGAGCACT-3’). Forward and reverse oligonucleotides were annealed using the annealing buffer. For annealing, sense and antisense oligonucleotides were mixed in equimolar concentration and annealing buffer (Appendix C). The mixture was heated for 5 minutes at 95°C that further cooled to 25°C in step down manner by decreasing per degree Celsius per 30 seconds. The annealed oligonucleotides were analyzed in 2% agarose gel. Further, [γ-<sup>32</sup>P]
Materials and Methods

labeled DNA probe was prepared by end labeling the annealed oligonucleotides using Polynucleotide kinase (Roche) using as per the Table 2.3.

Table 2.3: DNA probe labeling for EMSA.

<table>
<thead>
<tr>
<th>Component</th>
<th>Working concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annealed oligos</td>
<td>1 pmoles</td>
</tr>
<tr>
<td>5XPolynucleotide kinase buffer</td>
<td>1 X</td>
</tr>
<tr>
<td>[γ-32P] ATP</td>
<td>0.1 mCi</td>
</tr>
<tr>
<td>Polynucleotide kinase</td>
<td>10 U</td>
</tr>
<tr>
<td>Sterile double distilled H2O</td>
<td>Upto 20 µl</td>
</tr>
</tbody>
</table>

Incubate at 37°C for 30 min

The reaction was stopped using 1 mM EDTA. Labeled probe was purified using nucleotide removal kit (Qiagen) and eluted in Tris-EDTA buffer. Equal amounts of labeled probes were incubated with 5 µg – 15 µg of CTCF at 25°C for 20 min in EMSA binding buffer in a total volume of 20 µl. Control reactions were setup in the absence of CTCF. After the incubation, 6 X non-denaturing gel loading dye were added to the samples. The Protein:DNA complex and the free DNA probes were resolved by 5% non-denaturing polyacrylamide gel in running buffer (0.5 X TBE). Gels were then dried and subjected to autoradiography using Personal Molecular Imager (Bio-Rad).

2.8. In silico Modeling
We had checked for the DNA-Protein interaction at the atomic level with respect to time. We performed molecular dynamic simulation and quantum mechanics - molecular dynamic simulation for checking effect of variant in altering DNA binding of a protein transcription factor, CTCF.

2.8.1. HADDOCK Docking
The 3D-structure of CTCF protein (PDB ID: 2CT1) was retrieved from PDB (www.rcsb.org/). Water molecules were deleted manually. The protein structure was finally energy minimized using NOMAD-Ref server (http://lorentz.immstr.pasteur.fr
Materials and Methods

Active residues as these residues are involved in DNA binding (as observed from crystal structure) Figure 2.5. The starting structures for the docking were a B-form model of the double helix DNA fragments as follows:

Site_1: 5'-GTGGGCCTCG[C*]CCTGAGCACT-3',
Site_2: 5'- GTGGGCCTCG[T*]CCTGAGCACT-3',
Site_3: 5'- GTGGGCCTCG[C-CH3*]CCTGAGCACT-3'

These DNA templates were constructed using 3D-DART (http://haddock.Science.uu.nl/dna /dna.php). For the DNA fragment, bases 11–14 were selected as active residues based on the data as CCCTC is the binding site for CTCF factor.

Additional restraints to maintain base planarity and Watson Crick bonds were introduced for the DNA. During the rigid body energy minimization, 1000 structures were calculated, and the 200 best solutions based on the intermolecular energy were used for the semi-flexible, simulated annealing followed by an explicit water refinement. The solutions were clustered using a cutoff of 3.5-Å rmsd based on the pair wise backbone rmsd matrix. The semi-flexible annealing and the water refinement steps of HADDOCK were re-run with the best five structures of the lowest energy clusters (cutoffs 0.9 Å). The final 120 structures were clustered as described above, resulting in a single low energy cluster of 23 structures. The best 10 structures (rmsd 0.7 Å over backbone atoms) of this cluster were analyzed using standard HADDOCK protocols and were used to represent a model of the complex.

Figure 2.5: Enlarged secondary structure plot of 2CT1 with active residues marked.
2.8.2. Molecular Dynamics Simulation

MD simulations of wild type and mutant type DNA complex with CTCF protein in explicit water at 300 K were performed using the Gromacs 4.5.6 (Hess et al. 2008) and the Amber force field (Carletti et al. 2008). The TIP3 model was used for water molecules. The systems were equilibrated by 5000 steps of energy minimization, followed by a 250 ps MD simulation in the NVT ensemble, with harmonic restraints (20 kcalmol$^{-1}$ Å$^{-2}$) applied to the backbone atoms of the biomolecules. The harmonic restraints were gradually reduced to zero in a 750 ps MD simulation in the NPT ensemble. Production run in the NPT ensemble followed (200 ns for each system).

The temperature was maintained at 300K by Langevin dynamics with damping factor equal to 5 ps. Periodic boundary conditions were applied, and the pressure was kept at 1 atm by the Nose-Hover Langevin method, with an oscillation period of 200 fs and a damping time of 100 fs (Hoover et al.1985; Nosé et al. 1984). A smoothed cut-off (10–12Å) was used for the van de Waals interactions. Electrostatic forces were computed by the Particle Mesh Ewald algorithm with a maximum grid spacing of 1.0Å. Bonds with hydrogen atoms were restrained by the SETTLE algorithm (Hess et al. 1987; Kawata et al 2001) to use a time step of 2 femto second.