eQTL

Elucidating biological pathways bridging genotype and phenotype is one of the fundamental challenges in genetic and genomic research. Gene variants that control phenotypes are typically discovered by combined linkage analysis and molecular validation. However, these genotype-phenotype associations do not expose the underlying biological pathways through which gene variants operate on phenotype. We explore genomic and computational approaches to uncover the biological pathways through which genetic loci influence phenotypes and predict the effects of genetic and transcriptional variations/ interventions on gene expression and physiological/ behavioral phenotypes of individuals with different genetic background.

In last few decades, genome wide association studies (GWAS) in several species have led to the discovery of numerous loci for a range of complex diseases and many economically important traits in plants. mRNA expression data leads to the identification of expression quantative trait loci (eQTL), a genomic loci which is responsible to regulate the expression levels of various mRNA and proteins.

The trait value measured by mRNA or proteins is always a product of single gene with a specific chromosomal location. Expression QTL are empirically divided into two classes cis and trans. We identify cis-QTL region in which eQTL region is mapped to approximate location of their gene-of-origin i.e when both eQTL and gene position overlap, the eQTL is considered to be cis-regulated. While trans eQTL region are those regions which are far from the location of their gene-of-origin i.e if the eQTL and gene location are non-overlapping, eQTL is considered to be trans-acting. Any eQTL identified can be either cis regulated or trans regulated.

The polymorphism of the regulatory elements directly regulates the abudance of a gene transcript. The combination of whole genome-wide association studies with the measurement of global gene expression allows the systematic identification of eQTL. Xcelris has developed a pipeline where we simultaneously assay gene expression along with the genetic variation on a genome wide basis in large number of individuals along with statistical genetic methods which are used to map the genetic factors that determine the expression of many thousands of transcripts.

Global eQTL analysis in yeast, mice and humans have detected significant levels of eQTL traits that simultaneously regulate a large fraction of transcriptome. In yeasts, the complex inheritance of transcript levels has been revealed by detecting significant levels of nonadditive genetic variance, epistasis interactions and transgenic segregation. However, comparable studies in plants have yet to be reported. At Xcelris, we will provide eQTL analysis service both for the animal as well as plant samples.

Service type:

1. **eQTL Discovery**
2. **eQTL Bioinformatics**

eQTL Discovery:

Datasets to be provided for eQTL Discovery: DNA Genotype datasets, haplotype data and phenotype data.

Optional: Data can be generated by Xcelris (not in the scope of project and can be discussed mutually as per Xcleris catalogue) to obtain good quality and quantity suitable for small RNA experiment and will be charged separately.

Datasets to be generated by Xcelris for eQTL Discovery:

RNA-seq data

- 12-14GB of RNA seq data/sample on 2X300bp read length to be generated by Xcelris on MiSeq platform.
- SNP Discovery, SNP genotyping in population or germplasm as separate project
- Association studies (GWAS)

eQTL Bioinformatics

Datasets to be provided by customer for eQTL Bioinformatics: DNA genotype, haplotype data, phenotype data and 12-14 RNA-Seq data to be provided by customer.

Note: All types of sample sshould be transported in dry ice (-20^{o}C) containing cool packs to Xcelris Genomics, Ahmedabad, Gujarat, India.

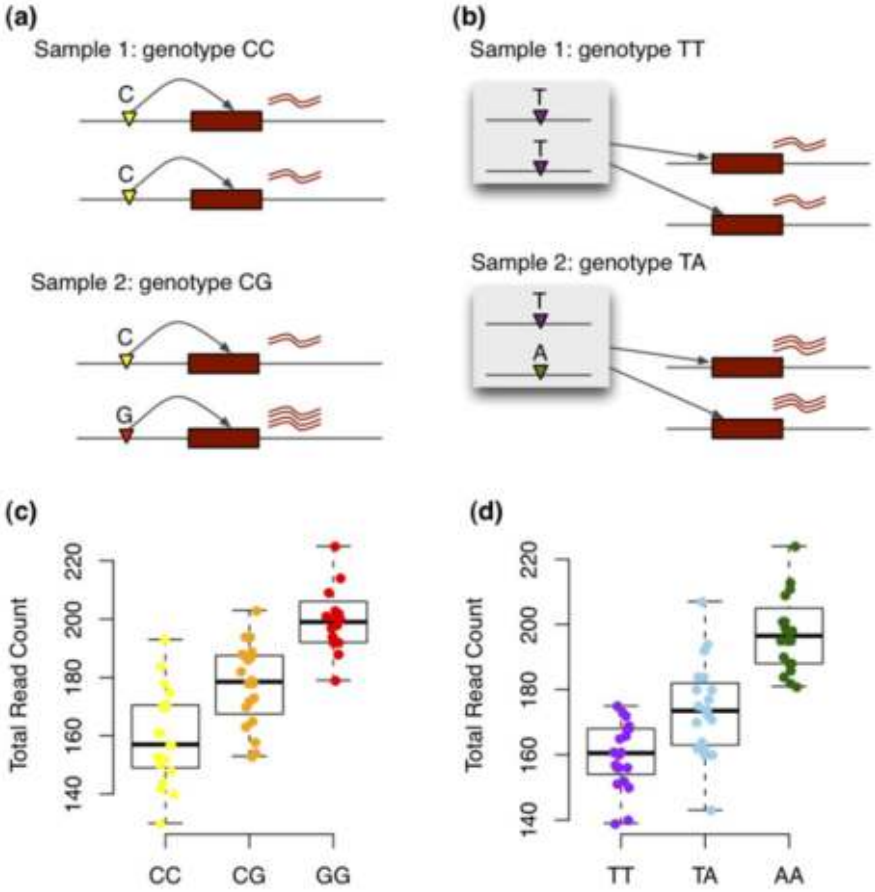


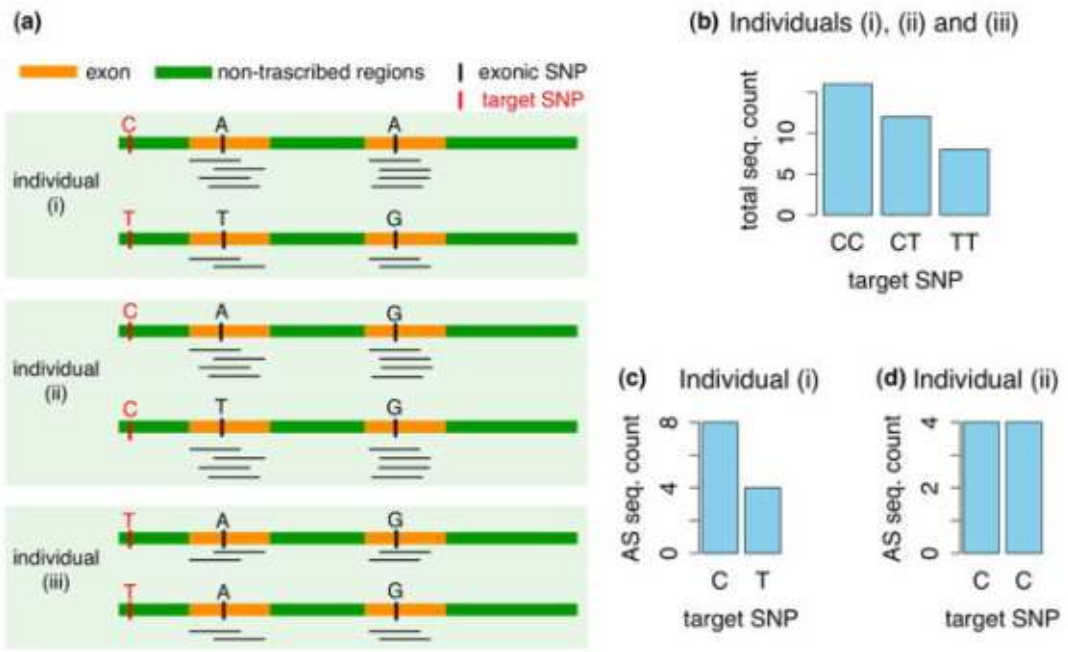
Figure 1:

(a) An example of a cis-eQTL in two samples. In Sample 2 where the target SNP (the SNP for which we test association) has a heterozygous genotype CG, the expressions of the two alleles are different.

(b) An example of a trans-eQTL in two samples. In Sample 2 where the target SNP has a heterozygous genotype TA, the expressions of the two alleles are the same.

(c) A simulated data for a cis-eQTL across 60 samples with 20 samples within each genotype class.

(d) A simulated data for a trans-eQTL across 60 samples with 20 samples within each genotype class.



Data Generation and work flow for eQTL:

Sample requirement for eQTL discovery:

- **Plant tissue:** (Seedling, leaves, stem, flower, fruits, graains etc.)
  Minimum 2-5 gm of tissue completely immersed in RNAlater should be provided and shipped in dry ice to Xcelris
- **Animal tissue:** (Blood, epidermal etc.)
  Minimum 2-5 gm of tissue completely immersed in RNAlater should be be provided and shipped in dry ice to Xcelris
  Note: All types of sample should be transported in dry ice(20OC) containing cool packs to Xcelris Genomics, Ahmedabad, Gujarat, India.
- **Isolated total RNA:**
  10-20 µg of total RNA should be provided along with RNA integrity number(RIN)>6. RNA must not be degraded and should be free from DNA contamination.
  Quality Control of RNA Sample: Samples will be subjected to both quantification, qualification and those with RIN>6 will be QC passed. Inclusion of low RIN value of the samples will be processed upon customer's confirmation.
  Optional: Xcelris will isolate total RNA from various parts of plant and animal tissues. Standardization is required for certain samples to obtain good quality and quantity suitable for small RNA experiments and will be charged separately.

RNA-seq Data generation:

- 12-14GB of RNA seq data/sample on 2X300bp read length to be generated by Xcelris on MiSeq platform.

Bioinformatics workflow for eQTL-Discovery/eQTL-Bioinformatics:

The implemantation of eQTL mapping using RNA-Seq can be divided into following steps:

1. **SNP/GWAS steady datas:**

- Phasing programs, such as BEAGLE or MACH will be used to impute the phase as well as to impute the genotype of a large set of SNPs that are phased against a referenced panel. RNA-seq reads will be aligned to the reference genome, call genotypes and then impute haplotypes using the genotype cells.

2. **RNA data processing:**

(a) Reference guided approach:

- The reads with low quality score or mapping ambiguity or low mapping quality will be removed.

- RNA-seq regions which may harbor more than one SNP or SNPs will be filtered.
- Mapping of the RNA-seq reads will be done by done approaches to the genome using Cufflinks and Scripture :
- Either reads from all individual are mapped to the same reference genome or they are mapped to the individual-specific haploid genomes that are constructed based on the phasing results.
- RNA-seq reads to the reference genome involve the detection of *de novo* exons and exons junction using TopHat, SliceMap, MapSlice, SplitSeek, QPALMA etc. Program.

(b) de novo assembly of RNA seq data using Velvet, AB ySS, CLC, trans AB ySS and Trinity will be used for denovo assembly of genes and annotation.

(c) RNA read counting:

- Total Read Count (TreC) per gene, per sample along with the number of allele-specific reads per allele of a gene, per sample will be calculated i.e if we have m genes and n samples, the counting TreC is a matrix of size mXn. The size of counting allele specific gene expression(ASE) will be mX2n.

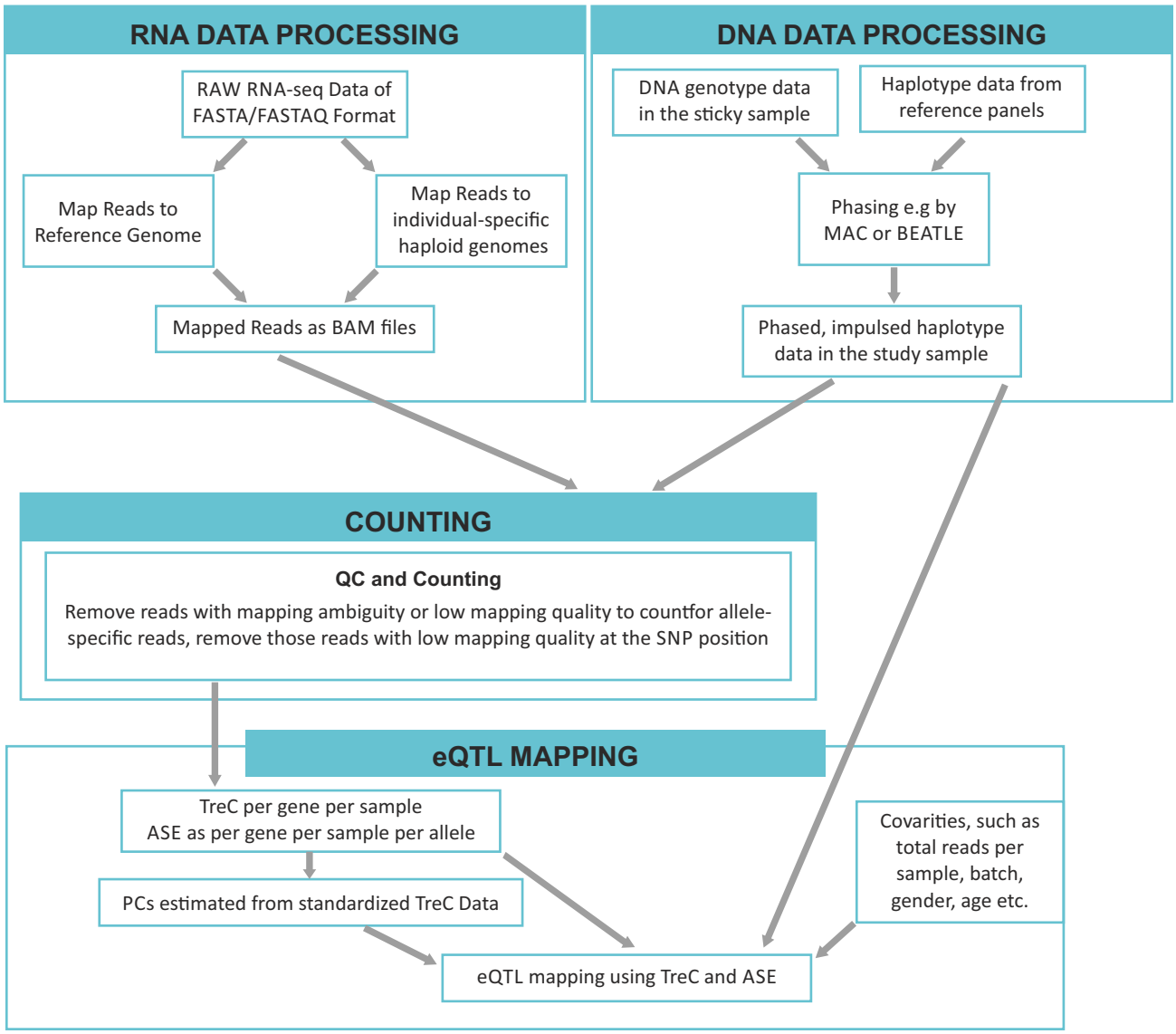


Figure 3: A workflow of eQTL mapping using RNA-seq data

3. eQTL mapping and Isoform Abundance Estimation:

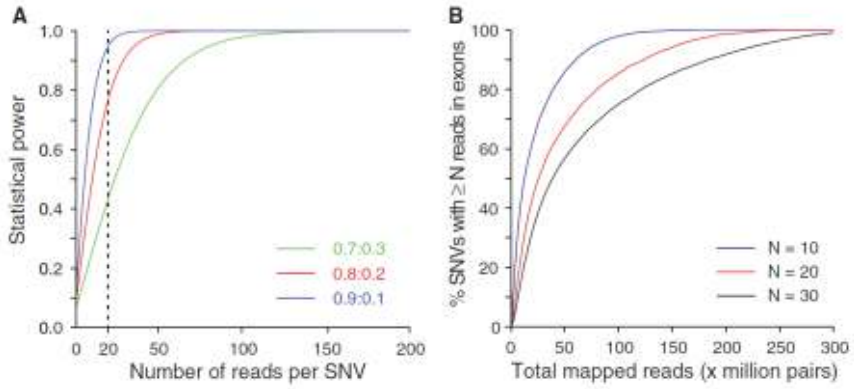
- The haplotype information provided by the client will be used to connect the alleles of the gene to the alleles of the target SNP. Principal Components Analysis (PCAs) and Principal Components(PCs) will be estimated by the TreCs of all genes of a sample normalized by the total number of reads of that sample.
- Isoform-specific eQTL mapping (splicing QTL mapping) will be done by dissecting the genetic basis of differential isoform usage using ALEXA-seq and NEUMA.
- The outcome of the above analysis will be represented as a statistical association between genetic markers located at specific regions in the genome and transcript abundance of the assayed gene. The eQTL plot significance of association will be derived recorded as the logarithm of odds(LOD) score or Likelihood Ratio Statistics (LRS), and plotted relative to each test position covered by the genetic markers across the genome.

Data Delivarables:

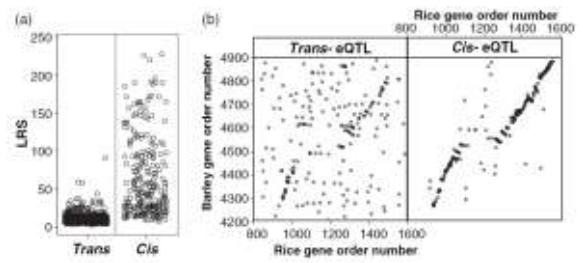
- 12-14GB RNA Seq data/ sample
- Fine mapping of eQTL regulating gene
- Genome wide association link
- Map position of e QTL and it's list
- eQTL biology replicates

(Reference: Sun W, Hu Y. (2013) eQTL Mapping Using RNA-seq Data. Stat Biosci 5(1), 198-219)
Gang L., et al. (2012)Identification of allele-specific alternative mRNA processing via transcriptome sequencing. Nucleic Acid Research 40(13)
Arnis D., et al. (2010) Expression quantitative trait loci analysis in plants. Plant Bio journal 5(1), 10-27

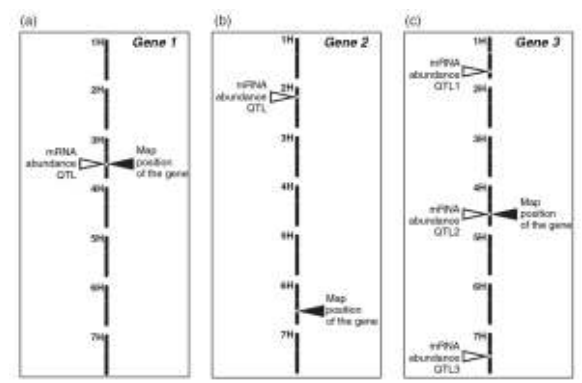
Statistical power and read coverage for ASE analysis



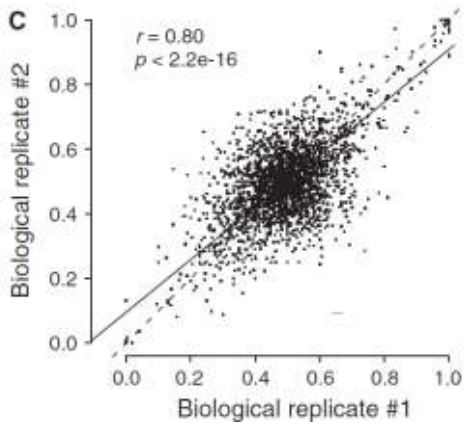
Validation of eQTL



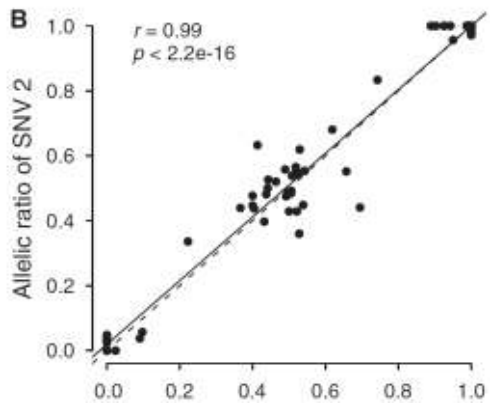
Gene expression regulation inferred by eQTL



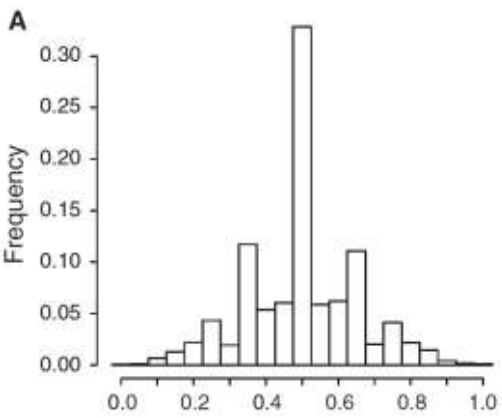
Alleleic ratio of SNVs in RNA-seq reads



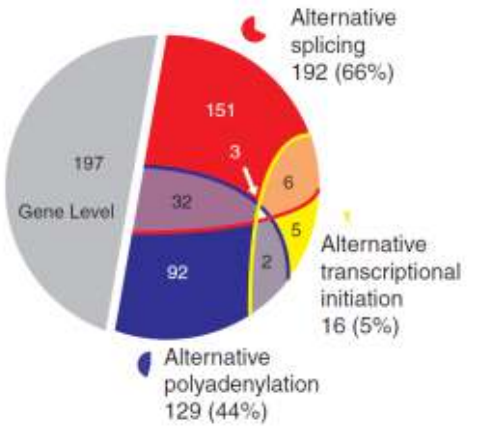
Allelic raio of SNV1



Allelic raio of SNV in RNA-seq reads



ASE pattern with genes



Summary of eQTL performed on different plant samples

Tissue used for array	No. genes/eQTL: Genes analysed	Genes mapped	Single eQTL	*cis*-eQTL	*cis*-eQTL (%)
Germinating embryos	15967	12933	5764		29–39
Ear leaf tissue	18805	6481	NA	NA	NA
5-week old stems	439	89	23	NA	NA
20-month old xylem	2608	1067	821	NA	NA
6-week old plants	22746	15664		5127	32%
Aerial parts of seedlings	24065	4066		1875	46%