FINAL REPORT - UGC Annexure-IX

Major Research Project

PROJECT TITLE

"Association Mapping for oil content in seeds of Jatropha curcas L. using EST derived Microsatellites Markers"

P.I. Dr. Neha T. Patel

Department of Biochemistry

Shree Manibhai Virani & Smt. Navalben Virani Science College

'Yogidham Gurukul', Kalawad Road, Rajkot - 360 005 (Gujarat)

12. Whether objectives were achieved (Give details):

OBJECTIVES:

During this project, 7 major objectives were achieved which is listed below:

1. Collection of *Jatropha curcas* germplasm from different geographical regions of India.
2. Measuring and recording phenotypic characteristics (the percentage oil content) in selected population group
3. Comparative analysis of different DNA extraction procedures in *Jatropha curcas* L.
4. Development of SSR primers from EST database of *Jatropha curcas*
5. Genotyping of collected samples with EST derived microsatellite marker.
6. To analyze the microsatellite generated molecular data for Assessment of Population structure and Kinship.
7. To develop molecular markers for identification of elite *J. curcas* genotypes having attribute of high oil content through Marker trait Association.

13. Achievements from the Project:

Achievements of the Project

- **Standardized and successfully implemented a novel modified version of Dellaporta *et al.* protocol for DNA isolation from *Jatopha curcas***: Comparative analysis of six different protocols was conducted; Dellaporta *et al.* displayed the highest resolution DNA bands. However, minor modifications of the protocol increased the yield as well as purity.
- **89 non-interrupted insilico validated EST-SSR primers have been designed**: Informative SSR markers from a large collection of EST database (42,483 ESTs) using EST database of *Jatropha curcas* were characterized. A total of 3682 ESTs contained SSRs; a total of 2236 primers was successfully designed and used for the validation of the amplification in *in silico* condition. From these validation studies, 93 primers did not contain hairpin loops, self-primers and cross primers. Out of which 4 were interrupted EST-SSR primers and the rest 89 were non-interrupted EST-SSR primers.
- **Genetic diversity study among 91 *Jatropha curcas* genotypes has been conducted**: Population genetic parameters were used to find diversity among individual accessions. The mean expected Heterozygosity/ gene diversity (He) was 0.820 indicating that, 82 % the probability, that two alleles randomly selected from an accession of *Jatropha* germplasm were genetically different. FST values ranging from 0.041 to 0.123 stated that the majority of the variance was found within a population rather than between populations.
- **Association mapping of oil content in *Jatropha curcas* L. using EST-SSR Markers to result in four significantly associated markers to the % oil content:** Four molecular markers, SSR 23, SSR 42, SSR 53 and SSR 56, out of 48 markers were found to be significantly linked to the trait of oil content among the *Jatropha curcas* accessions. The **significantly associated** markers were the ones with p-values less than 0.05 and were considered as 95% statistically significant. p-value is the probability of an association between a marker and a trait expected by chance.

14. Summary of the Findings (in 500 words):

SUMMARY

Jatropha curcas is a potential plant species with a variety of applications and enormous economic potential. Phenotyping all the *Jatropha curcas* accessions revealed that the seed oil content ranges from as low as 18% to as high as 42%. Since DNA isolation from leaf samples of *J. curcas* accessions proved to be difficult due the excessive phenolic and polysaccharide content, various DNA isolation procedures were evaluated to identify the most suitable procedure for this plant species. Comparative analysis of the methods of Doyle and Doyle, Jobes *et al.*, Dellaporta *et al.*, Sudheer *et al.*, Nalini *et al.*, Anna Maria *et al.*- 1 and Anna Maria *et al.*-2 for DNA isolation clearly showed that the method proposed by Dellaporta *et al*. after modifications in buffer to tissue ratio, amount of SDS and PVP used, provided good quality DNA from leaf tissue of *Jatropha curcas*.

From the 42,483 EST sequences of *J. curcas,* 3682 SSR sequences were identified. Based on the 3682 SSR-containing ESTs identified, a total of 2236 primers were successfully designed and used for the validation of the amplification in *in silico* condition. From this validation studies, 93 primers were such that which do not contain hairpin loops, self-primers and cross primers, out of which 4 were interrupted EST-SSR primers and the rest 89 were non-interrupted EST-SSR primers. The genotype data generated by extensive screening of the *J. curcas* accessions through EST based SSR markers were used to evaluate population genetic parameters to find diversity among individual accessions. The PIC values in the range of 0.0 to 0.9941 with an average of 0.821 reflect a uniformly high allelic diversity among the SSR markers. The mean expected Heterozygosity/ gene diversity (He) was 0.820 which indicates that, the probability, that two alleles randomly selected from an accession of *Jatropha* germplasm are genetically different, is 82%. FST values ranging from 0.041 to 0.123 state that the majority of the variance is found within a population rather than between populations. An analysis of the genotype data in STRUCTURE algorithm identified ten clusters/true populations/sub populations into which the samples from all predefined populations were categorized; this demonstrates the presence of admixture within populations.

Since *Jatropha curcas* populations being cultivated in India are genetically ill defined, Association Mapping was used to identify the markers associated to genes involved in oil biosynthesis. Association mapping using the MLM approach in TASSEL software specified four positive associations or correlation between markers, SSR 23, SSR42, SSR 53 and SSR 56, and the oil content. The BLAST2Go annotation of SSR 53 showed that it has putative identity to PAD4 that specifically share sequence similarity to triacyl glycerol lipases required for fatty acid

breakdown in oil biosynthesis pathway. Similarly, SSR 56 has been annotated to ICE1 (Inducer to CBF gene) protein whose mutation can reduce the expression of genes downstream to CBF. Hence, both are indisputably associated with the oil content. Furthermore, SSR-42 has been annotated to gene that belongs to PPPDE thiol peptidase protein family. It can be postulated due to their peptidase function that they play a role in activation of important enzymes in the oil biosynthesis pathway. While, SSR-23 has not been annotated to a particular family of proteins, however, it is conserved region of hypothetical protein according to BLAST 2Go. Thus, all the four markers are associated with oil content and can be further used for Marker Assisted Selection.

15. Contribution to the Society (Give details):

Jatropha curcas has shown enormous potential as a biodiesel crop. Despite its worldwide plantations, productivity of the plant is far too low to be commercialized as renewable source of fuel at a big scale. Hence the aim of our project was to find SSR markers that can select for high oil yielding varieties of *Jatropha curcas*; without in fact growing the seeds into fruits. The fruiting takes at least up to two years. Thus we selected association mapping where randomly selected natural population is employed, to avoid using a mapping population and the fact that association mapping incorporates all the meiotic events that have occurred in the history. Hence, there would have been more opportunities for recombination to have taken place over several generations, between many alleles. However, over time with proceeding generations, all the associations with other genes will break due to succeeding recombination, and the mutation will only be in Linkage Disequilibrium with alleles that are physically linked. Thus, two loci with significant linkage disequilibrium will most likely be inherited together even after many historical recombination events or will exist in proximity in successive generations. This report was a first step where we found four markers significantly associated with the oil content phenotypic trait in *Jatropha curcas* through their annotation to the currently existing proteins. High and low oil yielding profiles in terms of product size will be produced with commercialized high oil yielding *Jatropha* genotypes (work in progress). This information can further be used to produce a molecular marker kit; in order to test the *Jatropha* seeds for high oil yielding prior to farming. Ultimately, a cost-effective and time saving method for selection and, also novel high oil yielding *Jatropha* varieties can be commercialized to be available to laymen, farmers and industrialists. Consequently, an attempt to replace the pure *Jatropha* oil or blends with the current diesel at our petrol pumps in India.

16. Whether any Ph.D. Enrolled/Produced out of the Project: **N.A.**

17. No. of Publications out of the Project (Please attach): **01**

Dr. Neha T. Patel
Principal Investigator

Dr. A. U. Patel

Principal
Shree Manibahi Virani And
Smt. Navalben Virani Science College,
Rajkot.

Ms. Shivani Patel
Co-Investigator

CANDIDATE GENE MAPPING: APPROACH, METHODS AND SIGNIFICANCE

Shivani Patel, Nirali K Patel

Department of Biotechnology, Shree M & N Virani Science College, Rajkot, Gujarat, India 360 005
smpatel@vsc.edu.in

Abstract

Candidate gene is a gene with known or assumed function that may affect genetic control of a trait and thus, can be considered a 'candidate gene' for this trait. The Candidate gene associates a gene to its phenotypic trait. These quantitative traits responsible may be biomedical, economical, and even evolutionary important studies. The traditional candidate gene identification is tedious due to limited information of molecular marker and, also lack of computational tools and software. However, digital candidate gene approach makes candidate gene identification reliable and rapid due to available literature database and gene ontology database. The Candidate gene mapping is successfully conducted with the identification of molecular marker, linkage map construction and Quantitative trait locus mapping. The candidate gene approach is important for determination of associated genetic variant with phenotype.

Key words: candidate gene mapping, genome architecture, candidate gene, DigiCGA, software.

{**Citation:** Shivani Patel, Nirali K Patel. Candidate gene mapping: approach, methods and significance. American Journal of Research Communication, 2013, 1(8): 199-204} www.usa-journals.com, ISSN: 2325-4076.

INTRODUCTION

A candidate gene is located in a euchromatin region of chromosomes suspected of being involved in the expression of a trait. The candidate gene approach has proven extremely powerful for studying the genetic architecture of a complex trait. Candidate gene approach is powerful method to detect a quantitative trait locus (QTL). QTLs mapping is frequently used to identify genomic regions associated with a phenotypic trait of interest. All genes in the QTL are candidate loci for the status of quantitative trait. The essential step in candidate gene mapping is the identification of candidate gene from the genome. The identification of candidate gene can be done using either, traditional candidate gene approach or digital candidate gene approach [1]. Traditional candidate gene approaches cannot be relied on to identify all of the genes influencing a complex trait as limited information of molecular marker, and positional cloning is very laborious even it is influenced by genome size[2]. Digital candidate gene approach is widely used than traditional candidate gene approach as, it is rapid. Traditional candidate gene approach is on the basis of evidence rather than reason. Veracity of candidate gene identification is comparatively higher by digital candidate gene approach.

1.1 Candidate gene mapping

The candidate gene (CG) approach has proved to be extremely powerful for studying the genetic architecture of complex traits, as it is a far more effective and economical method for direct gene discovery [6]. The candidate gene mapping is important since candidate gene is the sequence that is known to affect the trait(s) of interest [7]. The explanation of candidate gene approach states that a major component of quantitative genetic variation of phenotype is caused by functional mutation of a putative gene [6]. A putative gene is a piece of DNA thought to be a gene based on sequence (ex. Open Reading Frame) whose functional gene product i.e. the expressed protein is unknown. The candidate gene mapping approach is one of the most important studies for the genetic control of a given quantitative trait. Quantitative trait is a continuously varying, measurable character, affected by the variation present in one to numerous genes in combination with environmental variation. The CG approach consists of three chronological steps. First, CGs are proposed based on molecular and physiological studies (functional CGs) or based on linkage data of the locus being characterized (all closely linked genes may be positional CGs). Second, a molecular polymorphism must be revealed to localize the CGs on a genetic linkage map to look for genetic linkage between the CG markers and the loci being characterized, or to calculate statistical correlations between CG polymorphisms and phenotypic variation in a set of genealogically unrelated individuals. It is important to notice that these two strategies are fundamentally identical and can be conducted together or successively [4]. Third, if map co-segregation and/or statistical correlation have been found, complementary experiments must be conducted to confirm the actual involvement of the CG in the trait variation. This is the validation step [5].

1.2 Positional candidate gene approach

Positional candidate gene approach involves integrated genome scan and candidate gene analyses, in which identification of candidate gene is mainly based on physical linkage [7]. This physical linkage maps provide the exact location of genes or genetic markers on chromosomes [8]. The Positional candidate gene identifies a gene within the vicinity of QTLs [9]. The Positional candidate gene approach works on two different pedestals such as sequence comparison (QTLs mapping) and expression data (microarray) The positional candidate gene approach has been reported in different fields including the classical examples of DGAT1 in cattle, GDF8 in sheep[10][11][12]. When applying the position-dependent strategy, it is difficult to prioritize functional candidates harboured in the targeted region, which is frequently scanned through the microsatellites markers [6].In positional candidate gene approach a sequence comparison gives a better success with smaller confidence interval about QTLs [14].

1.3 Functional candidate gene approach

The function candidate gene approach work on identification of traitassociated with expressed gene [9]. In functional candidate gene approach putative candidate genesare statistically detected from the genes controlling large components of inheritable gene expression variation. To date, some researchers began to consider or use this approach for seeking candidate genes in different fields. For instance, by using this strategy, functional candidate genes for “eye muscle area” in pigs were identified [15]. The genetic analysis of variation in gene expression would provide valuable models for studying complex and quantitative traits [16]. In general, important biological features of traits are directly reflected by transcript pattern, and quantitative traits are usually the consequence of the structure of genetic regulatory networks and parameters that control the dynamics of these networks [17]. The rationale of function-dependent strategy states that the genes responsible for the variation of gene expression process are also responsible for the variation of trait, and the candidate

gene governing the major genetic component of trait variation can be mined from the pattern of gene expression profiles. In fact, gene expression profiles are increasingly analyzed in the search for candidate genes [18].

1.4 Digital candidate gene approach.

The most advanced approach for identification of candidate gene is digital candidate gene approach, usually denoted by DigiCGA. DigiCGA is product of bioinformatics. DigiCGA also known as an *insilico* candidate gene approach or computer facilitated candidate gene approach. The completion of animal genome projects have revealed a multitude of potential avenues for identifying candidate genes in which digital approach is an important one that enables the systematic identification of genes underlying biological traits [19].Gene functional similarity search tool (GFSST) is a digital resource that makes it possible to identify candidates by certain principles, e.g., functional similarity. Functional similarity based on Gene Ontology (GO) annotation is used in diverse applications like gene clustering, gene expression data analysis, protein interaction, prediction and evaluation. [20].The Gene ontology is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species [33].The DigiCGA can be classified into three different approaches. The ontology based approach uses gene functional information from biological ontology sources available through internet. The computational based approach works on many statistical algorithms and computational methods, which includes data mining [21], Hidden Markova analysis [22], cluster analysis [23], and, kernel based data fusion analysis [24]. The integrated identification approach is integration of more than one method including web data based resources like literature based resources, biological ontology resources [25] and molecular interaction principles.

Table1 List of software and online tools use for digital candidate gene approach

NO	Name	Website
1	GeneSeeker	http://www.cmbi.ru.nl/GeneSeeker/
2	GFSST	http://gfsst.nci.nih.gov
3	Endeavour	http://www.esat.kuleuven.be/endeavour
4	G2D	http://www.ogic.ca/projects/g2d_2/
5	SUSPECTS	http://www.genetics.med.ed.ac.uk/suspects/
6	TOM	http://www.micrel.deis.unibo.it/~tom/
7	BioMercator	http://moulon.inra.fr/~bioinfo/BioMercator
8	FunMap	http://www.bioinformatics.polimi.it/GFINDer/
9	PROSPECTR	http://www.genetics.med.ed.ac.uk/prospectr/
10	QTL Mixer	http://qtl.pzr.unirostock.de/qtlmix.php
11	CoGenT++	http://cgg.ebi.ac.uk/cogentpp.html
12	KNN classifier	available on request: jianzxu@hotmail.com

STEPS FOR CANDIDATE GENE MAPPING

The development of expressed sequence tag (EST) markers, help in candidate and comparative gene mapping.

The various steps involved in candidate gene mapping are:

Step1: Collection of mapping population

Step2: Phenotypic measurements

Step3: Collection of genotypic data using restriction fragment length polymorphism (RFLP) and expressed sequence tag (EST)

Step4: Homologue detection

Step5: Linkage map construction and statistical analysis.The genetic linkage map can be constructed using JOIN MAP version 3.0 (Stam 1993) [28]. The Kosambi estimation method was used to convert recombination frequencies to map distances in centi Morgans (cM). The molecular map was drawn with Map chart version 2.1 (Voorrips 2002) [30]. The linkage map was constructed using MAPMAKER, version 3.0 (Lander et al. 1987) [31]. The initial scan for QTL was done with MAPMAKER/QTL 1.1 (Lincoln et al. 1992).

CONCLUSION

Classically, a link between a gene and a quantitative trait can be hypothesized based on linkage information [32]. Completion of genome sequences and improved bioinformatics could facilitate *in silico* cross-matching of candidate sequences with QTLs. The creation of more powerful bioinformatics tools for gene annotation could facilitate the choice of functional candidates among and outside the positional candidate genes [33]. The CG approach has been used with success in human genetics, animal genetics [34] and in plant genetics [36][4]. The candidate gene approach is applicable to traits related to the metabolism (enzyme activities and substrate levels) [35]. Candidate gene is used to determine a phenotype of genetic trait. The candidate gene mapping is possible through identification of candidate gene applying the candidate gene approaches. The combination of positional and functional candidate gene approaches represents a helpful prerequisite for cloning the candidate genes [9]. The candidate gene approach has been shown to efficiently characterize QTLsin plants [37].

REFERENCES

1. Johnson, A., O'Donnell, C. (2009)An open access database of genome-wide association results. BMC medical genetics10: 6
2. Janis S. Fisler, and Craig H. Warden, Mapping of Mouse Obesity Genes: A Generic Approach to a Complex Trait.
3. Tabor HK, Risch NJ, Myers RM, (2002)Candidate-gene approaches for studying complex genetic traits: practical considerations, Nat Rev Genet. 3: 391-397.
4. Angaji SA, (2009) QTL Mapping: A Few Key points, International Journal of Applied Research in Natural Products, Vol. 2(2), 1-3.
5. Pflieger, S., Lefebvre, V., Causse, M., (2001).The candidate gene approach in plant Genetics: a review. Mol. Breed. 7: 275-291.
6. Mengjin Zhu and Shuhong Zhao,(2007)Candidate Gene Identification Approach: Progress and Challenges, international Journal of Biological Sciences, 3 (7):420-427.
7. Brian Kinghorn, Positional cloning, candidate genes, synteny/comparative mapping, University of New England
8. Konstantin V. Krutovskii and David B. Neale,(2001)Forest Genomics for conserving adaptive genetic diversity, Food and Agriculture Organization of the United Nations,

9. Manfred Schwerin*et. al*,(2004)QTL mapping and mining functional candidate genes affecting health –the German ADR QTL Dairy Cattle Project, Animal Science Papers and Reports vol. 1, 95-100
10. Thaller G, Kuhn C, Winter A, et al.(2003)DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. Anim Genet.; 34: 354-357.
11. Grisart B, Coppieters W, Farnir F, et al.(2002)Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res.; 12: 222-231.
12. Johnson PL, McEwan JC, Dodds KG, et al., (2005) A directed search in the region of GDF8 for quantitative trait loci affecting carcass traits in Texel sheep. J Anim Sci.;83:
13. Ron M, Weller JI., (2007)From QTL to QTN identification in livestock - winning by points rather than knock-out: a review.Anim Genet.38:429-439.
14. Karen marshall, gene discovery, IAFA korea, april2006.
15. PonsuksiliBS, Wimmers K, Schmoll F, et al., (2002) Porcine ESTs detected by differential display representing possible candidates for the trait “eye muscle area”. J Animal Breed Genet(2002)
16. Cheung VG, Spielman RS,(2002) The genetics of variation in gene expression, Nat Genet; 32: 522-525
17. Frank SA, (2003)Genetic variation of polygenic characters and the evolution of genetic degeneracy, J Evol Biol.;16: 138-142.
18. Gibson G, Weir B, (2005) thequantitative genetics of transcription, Trends Genet.; 21:616-623.
19. Glazier AM, Nadeau JH, Aitman TJ,(2002)Finding genes that underlie complex traits, Science.; 298: 2345-2349.
20. Zhang P, Zhang J, Sheng H, et al, (2006)Gene functional similarity search tool (GFSST), BMC Bioinformatics; 7: 135.
21. Perez-Iratxeta C, Bork P, Andrade MA,(2002)Association of genes to genetically inherited diseases using data mining, Nature Genet.; 31: 316-319.
22. The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Research: D440–4:36.
23. Pellegrini-Calace M, Tramontano A,(2006)Identification of a novel putative mitogen-activated kinase cascade on human chromosome 21 by computational approaches, Bioinformatics.; 22:775-778.
24. Freudenberg J, Propping P,(2002)A similarity-based method for genome-wide prediction of disease-relevant human genes, Bioinformatics.,18 (Suppl 2):S110-115.
25. De Bie T, Tranchevent LC, van Oeffelen LM, et al, (2007)Kernel-based data fusion for gene prioritization, Bioinformatics; 23: i125-132.
26. Hristovski D,Peterlin B, Mitchell JA, et al, (2005)Using literature-based discovery to identify disease candidate genes, Int J Med Inform.; 74: 289-298.
27. Gion, J.-M., Rech, Ph., Grima-Pettenati, J., Verhaegen, D., and C. Plomion, (2000)Mapping candidate genes in Eucalyptus with emphasis on lignification genes, Molecular Breeding.; 6: 441–449.
28. Neale, D.B., M.M. Sewell, and G.R. Brown., (2002)Molecular dissection of the quantitative inheritance of wood property traits in loblolly pine. Annals of Forest Science (in press
29. Stam, P.,(1993)Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J. 3: 739–744.
30. Voorrips RE, (2002)MapChart: Software for the graphical presentationof linkage maps and QTLs.JHered 93:77–78

31. Lander, E. S., P. Green, J. Abrahamson, A. Barlow and M. J. Daly(1987)MAPMAKER: an interactive computer package for con,
32. Pflieger, S.et al., (2001)The candidate gene approach in plant genetics: a review. Mol.Breed. 7,275–291
33. Harhay, G.P. ndKeele, J.W., (2003) Positional candidate gene selection from livestock EST databases using Gene Ontology. Bioinformatics 19, 249–255.
34. Price, A H., 2006. Believe it or not, QTLs are accurate! Trends in Plant Science. 11(5):213-216.
35. Jean-Louis Prioul et al, (1999) From QTLs for enzyme activity to candidate genesin maize, Journal of Experimental Botany, 50,337,1281–1288,
36. Byrne, P. F., Mcmullen, M. D., 1996. Defining genes for agricultural traits: QTL analysis and the candidate gene approach. Probe. 7: 24-27.
37. Byrne PF, McMullen MD, Snooks ME, Musket TA, Theuri JM,. 1996. Quantitative traitloci and metabolic pathways: genetic control of the concentration ofmaysin, a corn earworm resistance factor, in maize silks. Proceedingsof the National Academy of Sciences, USA 93, 8820–8825.