Melting into thin air: A preliminary analysis of pig genome DNA marker distribution on chromosome 4 in relation to QTL locations.

Zhi-Liang Hu [1], Xiao-lin Wu [2], Daniel Gianola [2,3] and James M. Reecy [1]

1. Department of Animal Science, Iowa State University, Ames, IA, USA ; 2. Departments of Animal Sciences & Dairy Science, UW-Madison, Madison, WI, USA; 3. Department of Biostatistics and Medical Informatics, UW-Madison, Madison, WI, USA

Abstract

With the completion of the swine genome assembly, additional analyses shall be carried out to better understand it. Genome annotation analyses are underway to discover all possible coding genes, as there is a wealth of structural genome feature data from 20+ year's of genome research. Features include VNTR (microsatellites), RFLP, PCR-RFLP, STS such as cDNA probes, CpG islands, and more recently, SNPs. The anonymous DNA markers have played fundamental roles in the development of linkage maps, radiation hybrid (RH) maps, physical maps, and in the findings of hundreds of quantitative trait loci (QTL). While new technologies are moving genome analysis to a new stage in terms of sequencing, the architecture of quantitative trait loci still remains in "clouds and mist". QTL meta-analysis approaches can pool effect-size estimates or combine p-values from multiple studies. Non-parametric methods with a Dirichlet process prior help to effectively capture QTL variation inherent from all known data, to reveal significant "clusters" of QTL (QTR), which may contain a single QTL or several QTL that affect the same quantitative trait. In this study, we analyzed further the genome distribution of the known DNA markers in relation to several such QTR locations on chromosome 4, as a pilot study in an attempt to provide new interpretations to the QTL meta-analysis results. Our results are not only useful to guide the QTL information mining and to help understanding of genome organization in the light of animal phenotypes it codes for, and genomic architecture of QTL.

Materials and Methods

The data sources for the chromosomal locations of various genomic structural features are listed in Table 1. For the purpose of location correlation analysis, the number for the presence of each structural feature was counted in equally divided "bins" along the length of the chromosome. As the basic measuring unit for QTL is centimorgan (cM), we choose 1 Megabase (roughly correspond to 1 cM) as the bin size.

Table 1. Methods and sources of pig chromosome 4 feature data.

Structural Features	Abbreviations	Date source and methods
CpG island predictions (a)	CpG island (a)	Predicted with software "cpgi130.pl" by Takai and Jones (Ref. 1)
CpG island predictions (b)	CpG island (b)	Predicted by Ensembl (see Acknowledgement for details)
Linkage marker locations	Linkage markers	From Roslin institute (see Acknowledgement for details)
60K SNPs locations	SNPs	Provided by Martien Groenen of Wageningen University
Known gene locations	Known genes	Ensembl annotation, downloaded with BioMart query tools
Micro RNA predictions	miRNA	By blast prediction of known miRNA from MIRBASE (Ref. 2)
GO annotation of genes	GO terms	Ensembl annotation, downloaded with BioMart query tools
Quantitative Trait Loci	ADG, BF, IMF etc.	Animal QTLdb - http://www.animalgenome.org/QTLdb/pig.html

Results

1. The location distributions of predicted CpG islands, known genes, matched GO terms, linkage markers, SNPs and predicted miRNAs are plotted as shown in **Figure 1**. The selection of the structural features is inclusive, i.e., we include a feature as long as we may get the location information, and the locations are abundant enough to form a distribution.

2. Preliminary meta-analysis of major known QTL on pig chromosome 4 by simple counts for presence of each QTL at every centimorgan (cM) interval is plotted in **Figure 2**. The criteria to choose a QTL is that the QTL has to be reported by at least two publications (i.e. having supporting evidence from at least two different laboratories, or QTL tested with two different resource populations).

3. The GO annotations to known genes are analyzed for their functional hierarchy distributions (Ref. 8) with reference to GO slim, and the distribution patterns are compared between neighboring regions (roughly every 13 Megabases; **Figure 3**). It turned out the GO classes proportions are not dramatically different between the neighbors.

4. For the location distribution that showed a tendency for a pattern (e.g. "known genes" among others), an overlay plot was made (**Figure 4,** [a][b]) for trends of co-variations. Correlation coefficients (r^2) were calculated at equally spaced intervals and plotted for patterns (**Figure 4,** [c], [d]). The size of the intervals was adjusted as an experimental factor to look for potential pattern (**Figure 4,** [c] and [d]). Apparently no obvious pattern was observed at this time.

Although there is no obvious distribution patterns found, we believe we may find them as we improve the methods and include more data for analysis (see "Discussion"). The results will be useful for positional QTL information mining and to aid candidate gene discovery searches.

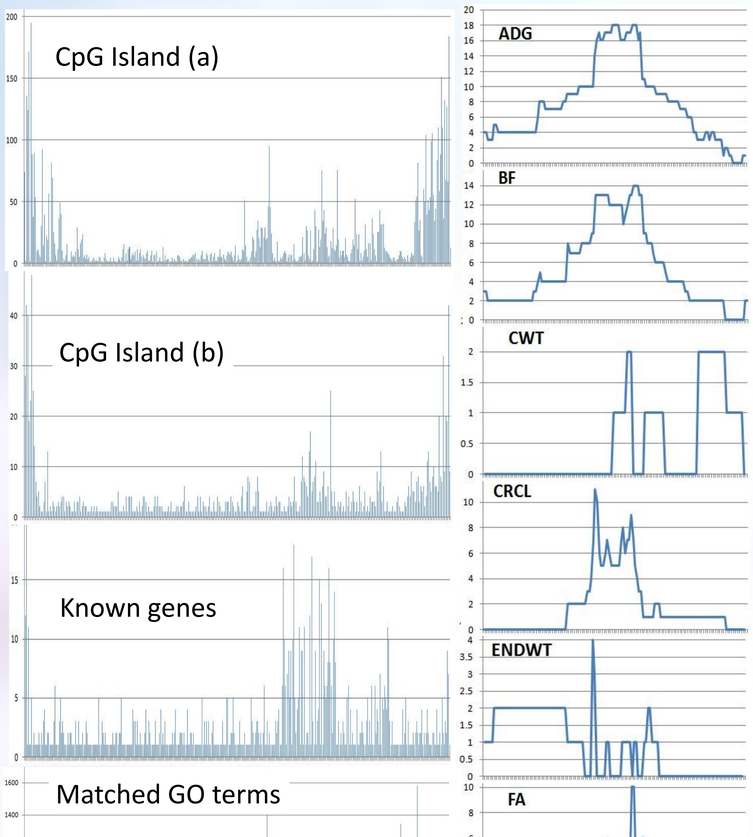

Figure 1
Distribution plots of predicted CpG islands, known genes, matched GO terms, linkage markers, SNPs and predicted microRNAs at bin sizes from 1Kb to 1Mb.

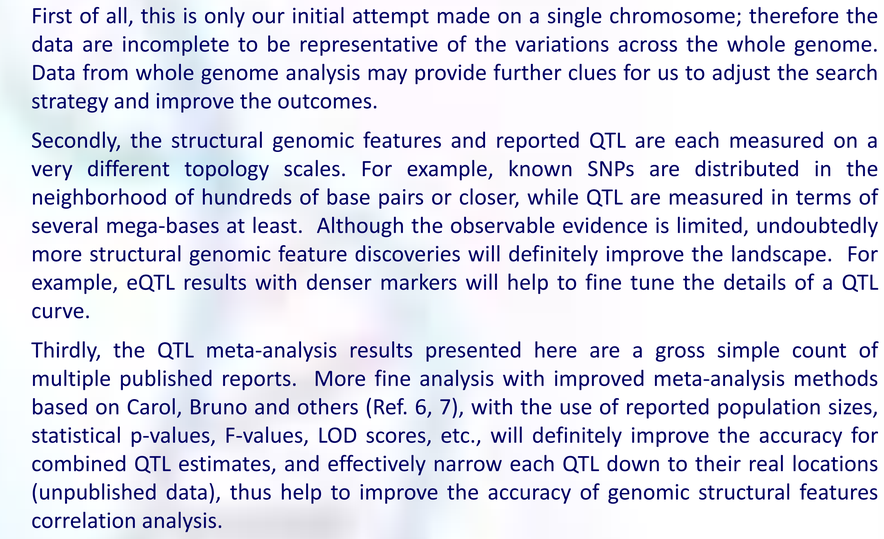

Figure 2
Preliminary meta-analysis of major known QTL on pig chromosome 4 by simple counts for presence of each QTL at every centimorgan (cM) intervals.

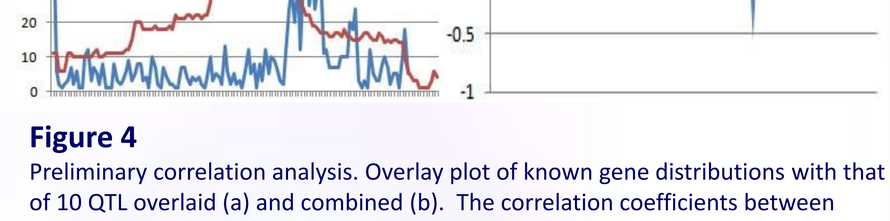

Figure 4
Preliminary correlation analysis. Overlay plot of known gene distributions with that of 10 QTL overlaid (a) and combined (b). The correlation coefficients between known gene densities and ADG (b) and BF (c), respectively, were also plotted for patterns.

Discussions

Known structural genomic features, such as CpG islands, SNPs, and microRNA, are believed to contribute to gene function variations (Ref. 3, 4, 5). Therefore, they must be associated with QTL variations in some ways. In this study, we attempted to find possible statistical correlations between the distribution of these structural genomic features and the locations of reported QTL up to date. Although the limited number of convincing correlations is not enough to support the assumptions, we believe this is only restricted by a number of factors:

First of all, this is only our initial attempt made on a single chromosome; therefore the data are incomplete to be representative of the variations across the whole genome. Data from whole genome analysis may provide further clues for us to adjust the search strategy and improve the outcomes.

Secondly, the structural genomic features and reported QTL are each measured on a very different topology scales. For example, known SNPs are distributed in the neighborhood of hundreds of base pairs or closer, while QTL are measured in terms of several mega-bases at least. Although the observable evidence is limited, undoubtedly more structural genomic feature discoveries will definitely improve the landscape. For example, eQTL results with denser markers will help to fine tune the details of a QTL curve.

Thirdly, the QTL meta-analysis results presented here are a gross simple count of multiple published reports. More fine analysis with improved meta-analysis methods based on Carol, Bruno and others (Ref. 6, 7), with the use of reported population sizes, statistical p-values, F-values, LOD scores, etc., will definitely improve the accuracy for combined QTL estimates, and effectively narrow each QTL down to their real locations (unpublished data), thus help to improve the accuracy of genomic structural features correlation analysis.

Finally, QTL for a trait are often seen distributed on multiple chromosomes. It would be more reasonable to evaluate such correlations across the whole spectrum of the genome, that may allow better fit of the assumptions.

References

1. Takai D and Jones PA (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22". Proc Natl Acad Sci U S A, 99(6):3740-5. (Web site: http://cpgislands.usc.edu/).
2. miRBase: the microRNA database. URL: http://www.mirbase.org/. Date download on Oct. 19, 2009.
3. Feil R, Berger F (2007). "Convergent evolution of genomic imprinting in plants and mammals". Trends Genet 23 (4): 192–199.
4. Michael Olivier (2004). "From SNPs to function: the effect of sequence variation on gene expression". Physiol. Genomics 16: 182-183
5. Muller Fabbri, Nicola Valeri and George A. Calin (2009). "MicroRNAs and genomic variations: from Proteus tricks to Prometheus gift". Carcinogenesis, 30(6):912-917.
6. Carol J. Etzel and Rudy Guerra (2002). "Meta-analysis of Genetic-Linkage Analysis of Quantitative-Trait Loci". Am. J. Hum. Genet. 71:56–65.
7. Bruno Goffinet and Sophie Gerber (2000). "Quantitative Trait Loci: A Meta-analysis". Genetics 155: 463–473.
8. Zhi-Liang Hu, Jie Bao and James Reecy (2008) "CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories". Online Journal of Bioinformatics. 9 (2):108-112.

Acknowledgements

Thanks are due to Steve Searle, Bert Overduin and Giulietta M. Spudich from Ensembl for their helps in obtaining Ensembl prediction of the CpG island data. The sharing of updated linkage marker locations on the most recent pig genome build by Trevor Paterson of Roslin Institute is appreciated. We also thank Shu-hong Zhao and Sheng-song Xie from Huazhong University, China for their kind efforts to locate predicted miRNA on chromosome 4 for this study (data not used).

Figure 3
GO Slim classification analysis of known genes on pig chromosome 4. The GO terms were counted by equally spaced intervals along the entire length of the chromosome.