Detection And Validation Of SNPs In Taiwan Country Chicken EST Libraries

E.-C. Lin[*], Y. H. Wang[*], C. X. Wang[*], C. W. Yeh[¶], C. H. Hsieh[¶], H. S. Sun[‡], M. C. Huang[§], S. T. Ding[*], S.-H. Chiou[§], B. R. Ou[†], C.F. Chen[§], Y. P. Lee[§] and W. T. K. Cheng[†]

Introduction

Single nucleotide polymorphism (SNP) is the basic and common variation in genomic sequence. The size of chicken genome is around 1.2×10^9 bp with 3.28×10^6 SNPs discovered (International Chicken Polymorphism Map Consortium, 2004). The international cooperative project used genomic DNA of broiler, layer, silky and red jungle fowl to identify those SNPs. Expressed sequence tag (EST) libraries have been used to detect abundant SNPs in exons (Kim, H., Schmidt C. J., Decker, K. S., and Emara, M. G. (2003)). Some of SNPs in exons change the corresponding amino acids in the related protein sequence, which are called non-synonymous SNPs.

The Taiwan Country chickens are the native breeds in Taiwan. There are 10 lines (5 male and 5 female) long-term selected with different criteria from an ancestral population in National Chung-Hsin University (Lee, Y. P., Yeh, L. T., and Huang, H. H., 1997). The Taiwan Country chickens are the most important meat-type breed with more than 60% market share in the local poultry market. It is important to maintain those lines for obtaining better heterosis and production efficiency. The information of genetic polymorphism is needed to achieve such goals in the local poultry industry. This study was designed to identify SNPs from several EST clone libraries of two major selection lines in National Chung-Hsin University.

Material and methods

Animals and Tissues for EST clone libraries. There were 4~6 hens randomly selected from the Taiwan Country chicken lines for the EST clone library construction: the line B for male with selection on growth rate and meat production; the line L2 for female with selection on egg production of 40 weeks of age. The description of the two lines selected from the same ancestral population for more than 20 generations were shown in Yang, K. T., Lin, C. Y., Huang, H. L. et al., 2008. Six different tissues (pituitary, liver, adipose, ovary, oviduct, and shell gland) related to egg-laying and muscle tissue from those hens during highly egg-

[*] National Taiwan University, Taipei, Taiwan
[¶] National Center for High-Performance Computing, Hsin-Chu, Taiwan
[‡] National Cheng-Kung University, Tainan, Taiwan
[§] National Chung-Hsing University, Taichung, Taiwan
[†] Tunghai University, Taichung, Taiwan

laying period were used. Totally, 42,404 EST clones from the 7 libraries were randomly selected for DNA sequencing using Applied Biosystems 3730 DNA Analyzer.

Sequence Analysis and SNP Detection. Those EST sequences were analyzed by using Phred (Ewing and Green, (1998); Ewing et al., (1998)) to base-calling and clean-up those low quality (QV$\leqq$ 30) bases and Phrap to assemble those remaining EST sequences. Those contigs obtained from Phrap were annotated and localized by Blastn (Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. (1997)) against NCBI (National Center for Biotechnology Information) NT and Genome databases. The contig with at least 5 ESTs were further predicted the possibility of any particular SNP happened $\geqq$ 15% of those overlapped ESTs by PolyPhred (Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997)). Those contig sequences with SNP predicted were Blastn against NCBI dbSNP database to make sure whether those SNPs found by PolyPhred are novel (not matched in dbSNP) or known (matched at least one SNP in dbSNP).

Validation of selected SNPs. The blood samples of 120 hens were collected equally from lines B and L2. The genomic DNA was extracted from those blood samples in our lab. Among those predicted SNPs with PolyPhred score $\geqq$ 98, there were 35 non-synonymous SNP (29 novel and 5 known and 1 negative control not detected with any SNP) sequences were selected for further validation using GenomeLab^{TM} SNPstream® SNP (Backman Coulter, CA) (Bell, P. A., Chaturvedi, S., Gelfand, C. A. et al. (2002)).

Results and discussion

There were 42,404 EST clones randomly selected for sequencing and 36,463 high quality sequences left after trimming those low quality bases by Phred. The results of assembling from those high quality sequences in the 7 Taiwan Country chicken EST libraries using Phrap were shown in Table 1. The total numbers of assembled contigs, sequences not aligned to other ESTs (Singlets), and sequences with problem during alignment and assembling were 3,455, 11,649, and 1,840, respectively. The high quality sequences ratio of 86.0% (36,463/42,404) was similar to the results in Kim, H., Schmidt C. J., Decker, K. S., and Emara, M. G. (2003).

Table 1: The number of sequences analyzed and contigs assembled using Phred and Phrap in different tissues of Taiwan Country chicken EST libraries

Tissues	Trace files	High quality	Contigs	Seq. in contigs	Singlets	Problem seq.
Pituitary gland	4,388	3,149	491	2,020	986	153
Ovary	7,146	6,348	1,443	4,264	1,771	346
Liver	6,686	6,033	1,065	4,448	1,513	355
Adipose tissue	7,907	6,682	1,532	4,862	1,461	353
Oviduct	5,961	5,202	702	3,546	1,504	276
Shell gland	7,087	6,569	1,368	2,985	3,559	311
Muscle	3,229	2,480	540	1,608	855	46
Total	42,404	36,463	3,455	23,733	11,649	1,840

The total number of SNPs predicted by PolyPhred was 1,107 with only 290 (26.2%) known and 817 (73.8%) still not matched in NCBI dbSNP (Table 2). With restriction of a contig with at least 5 ESTs and a SNP happened ≧ 15% of those overlapped ESTs in that contig, those SNPs predicted of very high Polyphred score (98~99) and moderately high (95~97) were 951 and 56 (91% of those total SNPs predicted).

Table 2: The number of predicted SNPs at different regions of genes in Taiwan Country chicken EST libraries[α]

	SNP predicted			PolyPhred score				
Locations	Known	Novel	Total	99	98	95~97	90~94	70~89
5' UTR	7	53	60	3	53	3	0	1
CDS	210	494	704	58	554	34	19	39
synonymous	177	368	545	47	435	28	9	26
non-synonymous	33	126	159	11	119	6	10	13
Intron	10	0	10	3	4	0	1	2
3' UTR	28	93	121	16	81	8	6	10
Pseudogene	0	5	5	0	3	1	0	1
Unknown	35	172	207	16	160	10	10	11
Subtotal	290	817	1,107	96	855	56	36	64

[α] The SNPs were Blastn against dbSNP to make sure whether those found by PolyPhred are novel (not matched in dbSNP) or known (matched at least one SNP in dbSNP).

The 35 non-synonymous SNP (29 novel and 5 known and 1 negative control) sequences of very high PolyPhred score were validated from 120 hens of lines B and L2. We found that 31 of true positive (88.6%), 3 of false positive (8.6%), and 1 of true negative (the negative control sequence, 2.8%). In addition, the five SNPs known in NCBI dbSNP were among the true positive (Table 3).

Conclusion

SNP prediction from EST sequences has been considered not as good as from genomic sequences in terms of quality. But the cost of obtaining abundant EST sequences is much cheaper than genome mapping project. In addition, EST clone library might be applied to those particular breeds and selection lines different from commercial and standard experimental ones. These results clearly show that the procedure of base-calling with certain QV, assembling contig and predicted SNPs with certain quality limitations might be greatly helpful for the total quality of prediction. Those high quality SNPs predicted from different breeds and lines will lead to a lot of important applications, i.e., phylogenetic analysis, functional studies, breeding programs, etc.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al*. (1997). *Nucleic Acids Res.* 25:3389-3402.

Bell, P. A., Chaturvedi, S., Gelfand, C. A. *et al.* (2002). *Biotechniques* Suppl:70-77.

Ewing, B. and Green, P. (1998). *Genome Res.* 8:186-194.

Ewing, B., Hillier, L., Wendl, M. C., and P. Green. (1998). *Genome Res.* 8:175-185.
International Chicken Polymorphism Map Consortium. (2004). *Nature* 432:717-722.
Kim, H., Schmidt C. J., Decker, K. S., and Emara, M. G. (2003). *Anim. Genet.* 34:249-254.
Lee, Y. P., Yeh, L. T., and Huang, H. H. (1997). *J. Chin. Soc. Anim. Sci.* 26:271-284.
Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997). *Nucleic Acids Res.* 25:2745-2751.
Yang, K. T., Lin, C. Y., Huang, H. L. *et al.*, (2008). *Mol. Cell. Probes* 22:47-54.

Table 3: The validation of predicted SNPs from Taiwan Country chicken EST libraries

SNP ID[α]	RefSeq accession no.	PolyPhred score	Substitution type	Validation[β]	dbSNP ID[γ]
NC	NM_204410	Null	Null	TN	---
SNP_0009	NM_205261	98	CTC^{Leu}/GTC^{Val}	TP	N/A
SNP_0037	NM_204290	98	AAT^{Asn}/AGT^{Ser}	TP	N/A
SNP_0058	NM_001031001	98	AGA^{Arg}/AGC^{Ser}	TP	rs15113610
SNP_0060	NM_001031001	99	CAA^{Gln}/GAA^{Glu}	TP	N/A
SNP_0073	NM_205060	98	AAA^{Lys}/GAA^{Glu}	TP	N/A
SNP_0093	NM_205463	98	AAT^{Asn}/AGT^{Ser}	TP	N/A
SNP_0107	NM_001044633	98	AAT^{Asn}/AGT^{Ser}	TP	N/A
SNP_0111	NM_001044633	98	CTT^{Leu}/GTT^{Val}	TP	rs15446954
SNP_0146	NM_204569	98	GCT^{Ala}/GTT^{Val}	TP	N/A
SNP_0355	NM_205168	98	ACG^{Thr}/ATG^{Met}	TP	N/A
SNP_0357	NM_205168	98	ACG^{Thr}/ATG^{Met}	TP	N/A
SNP_0363	NM_001031276	99	TCT^{Ser}/TGT^{Cys}	TP	N/A
SNP_0373	NM_204775	98	CTC^{Leu}/TTC^{Phe}	TP	N/A
SNP_0379	XM_417606	98	ATC^{Ile}/GTC^{Val}	TP	rs16180341
SNP_0401	XM_413860	98	AAG^{Lys}/GAG^{Glu}	TP	N/A
SNP_0543	NM_001030861	98	CCC^{Pro}/TCC^{Ser}	TP	rs15177583
SNP_0565	NM_207180	98	CTT^{Leu}/TTT^{Phe}	TP	N/A
SNP_0574	NM_204534	98	AAT^{Asn}/CAT^{His}	TP	N/A
SNP_0596	NM_204521	98	ATT^{Ile}/GTT^{Val}	TP	N/A
SNP_0634	XM_421513	98	ATG^{Met}/GTG^{Val}	TP	N/A
SNP_0638	XM_421513	98	AAG^{Lys}/GAG^{Glu}	TP	N/A
SNP_0645	XM_417119	98	CTG^{Leu}/GTG^{Val}	TP	N/A
SNP_0692	XM_001233402	98	CGC^{Arg}/TGC^{Cys}	TP	N/A
SNP_0705	NM_204541	98	ACT^{Thr}/CCT^{Pro}	TP	N/A
SNP_0751	XM_001232057	98	TTA^{Leu}/TTC^{Phe}	FP	N/A
SNP_0768	NM_205521	98	GCT^{Ala}/GGT^{Gly}	TP	N/A
SNP_0785	NM_001006561	98	GCG^{Ala}/GTG^{Val}	TP	N/A
SNP_0856	NM_204985	98	AAC^{Asn}/GAC^{Asp}	TP	N/A
SNP_0942	NM_001031368	99	ATA^{Ile}/GTA^{Val}	FP	N/A
SNP_0943	NM_001031368	98	ATA^{Val}/ATG^{Met}	FP	N/A
SNP_0949	NM_001006385	98	GAC^{Asp}/GAG^{Glu}	TP	N/A
SNP_1001	XM_415683	98	AAC^{Asn}/AGC^{Ser}	TP	N/A
SNP_1041	XM_426283	98	AAA^{Lys}/GAA^{Glu}	TP	N/A
SNP_1078	NM_001079717	98	ATG^{Met}/GTG^{Val}	TP	rs3137356

α. NC: the negative control.

β. TN: true negative; TP: true positive; FP: false positive.

γ. N/A: not available in the dbSNP.