doi:10.1111/j.1365-2052.2008.01809.x

The concordance test emerges as a powerful tool for identifying quantitative trait nucleotides: lessons from BTA6 milk yield QTL

E. Seroussi

Agricultural Research Organization, Institute of Animal Science, Bet-Dagan 50250, Israel

Summary

The lack of conventions for confirming the discovery of quantitative trait nucleotides in livestock was evidenced by the proposals of mutations in two different genes (*SPP1* and *ABCG2*) as the underlying functional mutation for a major quantitative trait locus (QTL) for milk concentration on bovine chromosome 6 (BTA6). Of these conflicting candidates, *SPP1* was excluded by follow-up studies and by the data described here. A simple test for concordance of the zygosity state between QTL segregation status and the candidate polymorphism was shown, in this case, to be a critical step towards establishing the proof. If a given sample effectively represents the genetic variation across the QTL region, haplotype-based concordance may further enhance the functionality and resolution power of this test, allowing identification of the causative gene.

Keywords ATP-binding cassette transporter, BCRP, OPN, OPN3907, osteopontin.

In a research focus article, de Koning (2006) gave an update on conflicting candidates for cattle quantitative trait locus (QTL). He focused on two reports (Cohen-Zinder *et al.* 2005; Schnabel *et al.* 2005) of variation in *SPP1* and *ABCG2*. These two genes have been proposed to underlie a major QTL with effects on protein and fat concentrations in milk; this QTL is located in the centre of chromosome 6 (BTA6). Comparing the arguments for each gene, de Koning suggested that both mutations are equally probable. In follow-up publications, these alternative quantitative trait nucleotides (QTNs) were examined. An evaluation in beef cattle populations revealed that although the *ABCG2* polymorphism was not segregating in these populations, the polymorphism in *SPP1* was associated with another QTL for post-weaning growth, and no effect on milk yield was observed (White *et al.* 2007). The *SPP1* polymorphism was also investigated for its effect on milk in buffalo and was not found to be significant (Tantia *et al.* 2008). Eventually, genetic support for a QTN in *ABCG2* affecting milk composition was presented in Norwegian Red cattle (Olsen *et al.* 2007).

The variation in *SPP1* (*OPN3907*) is an indel in a T tract, ~1240 bp upstream of the *SPP1* transcription-initiation site. Because *OPN3907* was proposed first as the QTN, de Koning put the onus on our group to demonstrate that it is not the underlying functional mutation, and to better specify our observation of no concordance of its segregation with the QTL genotypes in our population, which may have resulted from difficulty in obtaining good sequence reads. Indeed, sequencing PCR products with indels hampers base-calling, and the current tools (Seroussi *et al.* 2002; Seroussi & Seroussi 2007) only predict the actual sequence out of the resulting ambiguous sequence reads. Accurate genotyping of such a region would require subcloning of these products to distinguish the haplotypes present in each heterozygote. In several instances, *OPN3907* was sequenced using cloned DNA or homozygotes, revealing three distinct alleles. All cloned alleles, deposited in GenBank under accession numbers AJ871176, AC185945 and NW_931635, were of nine thymines (T_9) and were described as rare by Schnabel *et al.* (2005). These authors also sequenced an allele from a homozygote (T_{10}, AY878328). Sequencing of sire 3208 (Fig. 1c, lane 2) revealed a third allele with nine thymines followed by three adenines; only two adenines came after the T tract in the other alleles (Fig. 1a). Hence, this locus displayed length variation typical of a microsatellite with different numbers of repeats of either thymines or adenines. These alleles were designated *SPP1M1*, *SPP1M2* and *SPP1M3* respectively (Fig. 1a). Sequencing of heterozygotes resulted in ambiguous sequence reads, which were traced as follows (Fig. 1c): *SPP1M1* and *SPP1M2* (lanes 3, 5); *SPP1M2* and *SPP1M3* (lanes 4, 6). Following this scheme, a sample of genotypes is presented (Fig. 1b). Although the status of the *ABCG2* mutation was in concordance with the QTL status, concordance was not observed with either the length of the T tract or the allele status of the *SPP1*

Address for correspondence

E. Seroussi, Institute of Animal Sciences, A.R.O., The Volcani Center, Bet Dagan 50250, Israel.
E-mail: seroussi@agri.huji.ac.il

Accepted for publication 20 August 2008

AGE 1809 B
Journal Name Manuscript No.
Dispatch: 16.10.08 Journal: AGE CE: Anusha
Author Received: No. of pages: 5 PE: Prasanna

COLOUR

(a)

Allele	Source	Lane	sequence
SPP1M1(T_9)	NW_931635	1	AAGTTTTTTTTT-AA-TATAGAATTTTTATTTTCTT
SPP1M2(T_{10})	AY878328		AAGTTTTTTTTTTAA-TATAGAATTTTTATTTTCTT
SPP1M3(T_9)	Sire 3208	2	AAGTTTTTTTTT-AAATATAGAATTTTTATTTTCTT

(b)

Sire	Birth year	% prot.	*ABCG2* exon14	*OPN3907* allele	*SPP1M* allele	QTL status
3208	1988	3.11	11	T_9/T_9	*M3/M3*	N
3241	1989	3.09	11	T_9/T_{10}	*M1/M2*	N
3089	1984	3.08	11	T_9/T_{10}	*M3/M1*	N
3212	1988	3.05	11	T_9/T_{10}	*M3/M1*	N
2283	1985	3.04	11	T_9/T_{10}	*M3/M1*	N
2357	1987	3.02	11	T_9/T_{10}	*M3/M1*	N
3070	1983	3.01	11	T_9/T_{10}	*M3/M2*	N
9 [3]	1974	3.11	12	T_9/T_{10}	*M1/M2*	Y
3099	1985	2.89	12	T_9/T_{10}	*M1/M2*	Y
2278	1985	2.89	12	T_9/T_{10}	*M1/M2*	Y
2182	1982	3.04	22	T_9/T_9	*M1/M1*	ND
2227	1983	3.03	22	T_9/T_9	*M1/M1*	ND
3573	1994	2.92	22	T_9/T_9	*M1/M1*	ND
3396	1992	2.91	22	T_9/T_9	*M1/M1*	ND
3094	1984	2.91	22	T_9/T_9	*M1/M1*	ND
2176	1982	2.81	22	T_9/T_{10}	*M1/M2*	*ND*

(c)

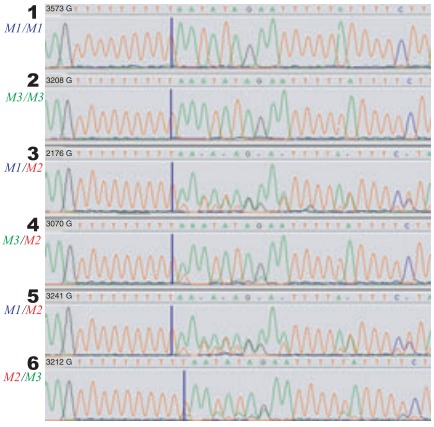

Figure 1 The genotypes of the *SPP1* microsatellite are not in concordance with the status of the BTA6 quantitative trait locus (QTL). Bovine genomic DNA was PCR-amplified and sequenced using primers F (5′-AGATCCCACATGCACCTAGC-3′) and R (5′-GCATGTCTCTGAAGGACTGG-3′), which were designed upstream of the *SPP1* transcription-initiation site. The three observed alleles were annotated as *SPP1M1*, *SPP1M2* and *SPP1M3* (blue, red and green respectively). Following the annotation of *OPN3907* (Schnabel *et al.* 2005), the lengths of their T tracts are presented (9T and 10T in orange and purple respectively). (a) The allele names, the source sequence (GenBank accession or genomic DNA), the associated lane in panel C and the nucleotide sequence are given. (b) A sample of the sires used by Cohen-Zinder *et al.* (2005) followed by their year of birth, their evaluation of milk-protein concentration as recorded by The Israeli Cattle Breeders Association (http://www.icba.org.il/cgi-bin/bulls/he/bl_main.htm), their corresponding genotype including the *ABCG2* plus (tyrosine-encoding) and minus (serine-encoding) alleles (1 and 2 respectively) and their QTL status: non-segregating (N), segregating (Y) and not determined (ND). This table is sorted first by *ABCG2* genotype and then by protein percentage. (c) Chromatograms are shown with their matching genotypes of sires 3573, 3208, 2176, 3070, 3241 and 3212 (lanes 1–6 respectively). The raw trace files are downloadable at http://cowry.agri.huji.ac.il/SPP1M. Note that traces interpreted as combination of *SPP1M1* and *SPP1M2* were labelled with either *SPP1M1*/*SPP1M2* or *SPP1M2*/*SPP1M1* according to the dominant base at the end of the T tract. In case of imbalance within the PCR amplification preferentially amplifying one of the alleles, the bases A or T would dominate this position according to the preferred allele (*SPP1M1* or *SPP1M2* respectively).

microsatellite. For example, the traces of the three sires segregating for the QTL were all of the type *SPP1M1*/*SPP1M2* (Fig.1c, lane 3) and were indistinguishable from that of the non-segregating sire 3241 (Fig. 1c, lane 5). This proves that of the two suggested causative mutations, only the variation in *ABCG2* is not excluded for the BTA6 QTL. Note that the entire *SPP1* gene was sequenced in two of our heterozygous sires (Cohen-Zinder *et al.* 2005) revealing all possible sequence polymorphisms in this gene for these sires. We encountered three heterozygous variations, none of which was concordant with the QTL genotypes, thereby ruling out *SPP1* as the causative gene for the QTL.

Henceforth, the *ABCG2* allele capable of encoding tyrosine, which was the more frequent allele in the population, will be denoted the plus QTL allele. The minus *ABCG2* allele (encoding serine) increases milk yield, and thus decreases fat and protein concentration. The sequence of a sire homozygous for the minus QTL haplotype was associated with *SPP1M1* (T_9) (Fig. 1c, lane 1). De Koning suggests that in the evaluated population there might be perfect linkage disequilibrium (LD) between these *SPP1* and *ABCG2* alleles. However, the results shown in Fig. 1 indicate that this is not the case. Sire 2176, which has one of the lowest protein percentages ever recorded in Israel, is homozygous for the *ABCG2* allele encoding serine but heterozygous for *SPP1M* (Fig.1c, lane 3). Moreover, sire 3208 had the highest protein percentages in our study and is therefore unlikely to be homozygous for the minus QTL allele. This sire was indeed homozygous for the plus allele, but also for the *SPP1M3* (T_9) minus allele. Genotyping this *SPP1* polymorphism using the scheme described here should better clarify its association with post-weaning growth (White *et al.* 2007). Hence, further genotyping of this locus is needed to effectively apply this observation to beef cattle.

De Koning's article prompted other groups to corroborate the identity of the BTA6 QTN. In this respect, note that sire 9 (Fig. 1b), which segregates for the QTL, has been heavily used in global breeding programs. This sire had a high protein percentage (3.11%) despite carrying the minus QTL allele, which has a relatively large effect on protein percentage (−0.21%, Cohen-Zinder *et al.* 2005). Still, it is possible that sire 9 was also a carrier of positive alleles at other genetic loci that have yet to be characterized, and the combination of these alleles with the BTA6 QTL allele (which increases milk yield) has made this an exceptionally popular bull. The need for better standards for reviewing QTN identification is evident from de Koning's research

focus. De Koning notes that Schnabel *et al.* (2005) tested the concordance of the segregation status of single nucleotide polymorphisms (SNPs) in their entire experimental population but, 'curiously', not that of *OPN3907*.

A simple test for concordance of the zygosity state between the QTL segregation status and the candidate polymorphism has emerged as a powerful tool for identifying the functional mutation underlying the QTL. In the search for QTN among the abundance of known SNPs, this cost-effective approach enables the genotyping of only a few sires with known QTL status, the latter having been determined beyond doubt using proper statistical analyses, such as the daughter design. When a concordant SNP is encountered, further genotyping may also be applied to determine the concordance of phase. In a sample of 18 sires with known BTA6 QTL status from the US and Israeli Holstein cattle, only three were heterozygous for the QTL (Cohen-Zinder *et al.* 2005). In the Norwegian Red, two and 16 sires were heterozygous and homozygous respectively (Olsen *et al.* 2007; Olsen, personal communication). The probability for a single match by chance alone between zygosity states and QTL status is therefore equal to the frequency of occurrences of heterozygotes ($N_{\text{hetero}} = 5$; $P = 0.139$) and homozygotes ($N_{\text{homo}} = 31$; $1-P = 0.861$). Hence, the probability for overall zygosity concordance by chance can be estimated to be as low as 5×10^{-7} using the formula $P^{\text{Nhetero}} \times (1-P)^{\text{Nhomo}}$ (Olsen *et al.* 2007). On one hand, this formula does not take into account phase concordance with the QTL allele; on the other, it problematically assumes that the sires are not related. Phase determination via the construction of haplotypes for the sampled sires provides a better concordance test that is based on chromosome haplotypes.

Haplotypes within the critical chromosomal region of the BTA6 QTL were inferred from the genotype data of the sire pedigree (Baruch *et al.* 2006; Olsen *et al.* 2007). Haplotype data make it possible to estimate the degree of relatedness among the chromosomes represented within these samples and, by consulting the bovine haplotype-block map (Khatkar *et al.* 2007), these data can help verify that the sample of analysed chromosomes is indeed representative in Holstein-Friesians.

In the two referred studies, all homozygous sires carried the common allele (Cohen-Zinder *et al.* 2005; Olsen, personal communication), bringing the frequency of chromosomes carrying the rare allele in these samples to 0.07, which is also the estimated probability (P) of encountering a chromosome by chance with the rare allele that is also a plus QTL chromosome. It is likely that all of these homozygous sires were likewise homozygous for the common plus QTL allele. The modified grand-daughter design (MGD) can be used to determine the QTL genotype of grandsires homozygous for the QTL (Weller *et al.* 2002). Hence, verifying this QTL status in the homozygotes would have been possible if MGD had been used. Assuming this QTL status had been determined, the overall probability for haplotype concordance by chance in these samples (number of concordant plus chromosomes, $N_{+\text{chr}} = 5$; number of concordant minus chromosomes, $N_{-\text{chr}} = 67$) could have been estimated by a similar formula, $P^{N+\text{chr}} \times (1-P)^{N-\text{chr}}$, to have dropped to 1.3×10^{-8}. However, phase concordance was verified for the minus haplotypes and only partially for the plus haplotypes using MGD (Weller *et al.* 2002). Moreover, some of the chromosomes in these samples were identical along 6 Mb, as noted for the two minus haplotypes derived from the Red Norwegian population. These haplotypes belonged to two sires that were paternal halfsibs (Olsen *et al.* 2007) and therefore should not be regarded as independent events for this test.

The LD map of BTA6 revealed extensive LD that may limit the level of resolution achievable in fine mapping (Khatkar *et al.* 2006). Olsen *et al.* (2007) noted the potential existence of another variant that is concordant with QTL status all along the 6-Mb haplotype shared between their two heterozygous sires. They raised the possibility that such variation could be the true causal mutation and that the QTN suggested in *ABCG2* was merely a marker in perfect LD with it. Our group had similar reservations (Cohen-Zinder *et al.* 2005), although the common plus QTL haplotype shared among the three heterozygotes was quite limited, around 0.6 Mb between the markers in the *MED28* and *ABCG2* genes (Baruch *et al.* 2006). Such reservations may underestimate the true power of the concordance test. The two closest *SPP1* SNPs genotyped by our group were 3.5 kb apart. Despite residing within the same haplotype block (Block 2, Table 1), these SNPs revealed all four possible allele combinations in the chromosome sample used for our concordance test. Assuming a minimal mutation rate, at least one historical recombination would be required within this short interval to obtain all combinations. A resolution power of such magnitude would indeed allow pinpointing the gene underlying the QTL. Olsen *et al.* (2007) also dismissed their hesitations by indicating that LD in the region is generally quite low and that perfect LD appears to be rare in their population. Moreover, the suggested *ABCG2* QTN is located between and not within haplotype blocks (shaded in dark or light grey, Table 1), which further reduces the odds of the occurrence of other variants displaying complete LD with this QTN.

The concordance test has its drawbacks. It will fail if different mutations in the underlying gene produce the same phenotype. However, considering the limited variation in dairy cattle, such cases are expected to be rare. It is also difficult to estimate the resolution power of this test because it depends on unrecorded historical recombination events. Nevertheless, it is possible to confirm that the chromosomes represented within the sample are unrelated by determining their haplotypes. Full haplotype concordance of a representative sample of chromosomes is likely to identify the QTN or at least point to the causative gene.

Table 1 Haplotype blocks near *ABCG2*.

Position build 3.1	Gene build 3.1	Khatkar SNP #	Cohen-Zinder SNP #	Olsen SNP #	Nucleotide variant	Block no.[1]
33473299	*PPM1K*		PPM1K ex2			ND
33473593	*PPM1K*	463001			A>G	31
33482293	*PPM1K*			PPM1K_309	A>C	ND
33489402	*PPM1K*	465459			A>G	31
33514410	Intergenic			BZ916464_39	G>T	ND
33514516	Intergenic			BZ916464_145	A>G	ND
33514682	Intergenic			BZ916464_311	C>T	ND
33514775	Intergenic			BZ916464_404	A>G	ND
33514831	Intergenic			BZ916464_460	C>G	ND
33583282	*ABCG2*	464405			C>G	31
33583352	*ABCG2*	465201			C>G	31
33583613	*ABCG2*		ABCG2(1) int3		A>T	ND
33584056	*ABCG2*	466263			C>T	31
33584124	*ABCG2*	466354			C>T	31
33584316	*ABCG2*	464963			C>T	31
33598812	*ABCG2*	463364			A>G	31
33616733	*ABCG2* QTN	ABCG2(2) ex14	ABCG2_49	A>C		
33617353	*ABCG2*			ABCG2_256	A>G	
33631457	Intergenic			AAFC02144624_75784	A>G	1
33631732	Intergenic	348220		AAFC02144624_03129	A>G	1
33632007	Intergenic	353109		PKD2_746 AAFC02144624_03128	A>G	1&2
33632945	Intergenic			PKD2_1175	Indel A	2
33635701	LOC530393			PKD2_1451	A>G	2
33635803	LOC530393			PKD2_1349	A>T	2
33636502	LOC530393			PKD2_650	C>T	2
33636799	LOC530393			PKD2_353	C>T	2
33637745	LOC530393			PKD2_349	G>T	2
33638005	LOC530393			PKD2_611	A>G	2
33638006	LOC530393			PKD2_610	A>T	2
33640579	LOC530393			PKD2_383	A>G	2
33644751	LOC530393			PKD2_901	A>G	2
33647488	LOC530393			PKD2_377	A>T	2
33647558	LOC530393			PKD2_447	A>G	2
33648352	LOC530393			PKD2_1241	C>T	2
33649870	LOC530393			PKD2_2759	C>T	2
33650721	LOC530393			PKD2_3610	C>T	2
33651020	LOC530393			PKD2_3909	A>T	2
33651301	LOC530393			PKD2_97141	A>G	2
33659174	LOC530393			PKD2_1013	A>G	2
33659234	LOC530393			PKD_953	C>T	2
33659590	Intergenic			PKD2_597	C>T	2
33693618	Intergenic		PKD2		Indel A	ND
33710689	*SPP1*	465810	SPP1(2) ex7		A>C	ND
33712386	*SPP1*		SPP1(1) int5		A>G	ND
33718525	Intergenic		OPN3907	OPN_3907	Indel T	2

[1]SNPs within blocks are shaded in grey; the block numbers 31 (dark grey), 1 (light grey) or 2 (dark gray) follow those from Khatkar *et al.* (2007) and Olsen *et al.* (2007). SNPs for which block status was not determined (ND) by these studies are also indicated.
QTN, quantitative trait nucleotide; SNP, single nucleotide polymorphism.

Acknowledgements

We acknowledge the contribution from the Agricultural Research Organization, Institute of Animal Science, Bet Dagan, Israel, No. 510/07. This research was supported by the Israel Milk Marketing Board.

References

Baruch E., Weller J.I., Cohen-Zinder M., Ron M. & Seroussi E. (2006) Efficient inference of haplotypes from genotypes on a large animal pedigree. *Genetics* **172,** 1757–65.

Cohen-Zinder M., Seroussi E., Larkin D.M. *et al.* (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15,** 936–44.

Khatkar M.S., Collins A., Cavanagh J.A. *et al.* (2006) A first-generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics* **174,** 79–85.

Khatkar M.S., Zenger K.R., Hobbs M. *et al.* (2007) A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in holstein-friesian cattle. *Genetics* **176,** 763–72.

de Koning D.J. (2006) Conflicting candidates for cattle QTLs. *Trends in Genetics* **22,** 301–5.

Olsen H.G., Nilsen H., Hayes B., Berg P.R., Svendsen M., Lien S. & Meuwissen T. (2007) Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. *BMC Genetics* **8,** 32.

Schnabel R.D., Kim J.J., Ashwell M.S., Sonstegard T.S., Van Tassell C.P., Connor E.E. & Taylor J.F. (2005) Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 6896–901.

Seroussi Y. & Seroussi E. (2007) TraceHaplotyper: using direct sequencing to determine the phase of an indel followed by biallelic SNPs. *Biotechniques* **43,** 452–3.

Seroussi E., Ron M. & Kedra D. (2002) ShiftDetector: detection of shift mutations. *Bioinformatics* **18,** 1137–8.

Tantia M.S., Mishra B., Bharani Kumar S.T., Mishra B.P., Kataria R.S., Mukesh M. & Vijh R.K. (2008) Characterization of Osteopontin gene of *Bubalus bubalis*. *Animal* **7,** 987–90.

Weller J.I., Weller H., Kliger D. & Ron M. (2002) Estimation of quantitative trait locus allele frequency via a modified granddaughter design. *Genetics* **162,** 841–9.

White S.N., Casas E., Allan M.F., Keele J.W., Snelling W.M., Wheeler T.L., Shackelford S.D., Koohmaraie M. & Smith T.P. (2007) Evaluation in beef cattle of six deoxyribonucleic acid markers developed for dairy traits reveals an osteopontin polymorphism associated with postweaning growth. *Journal of Animal Science* **85,** 1–10.

Author Query Form

Journal: AGE

Article: 1809

Dear Author,
During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Many thanks for your assistance.

Query reference	Query	Remarks
Q1	**Au: Please define: BCRP and OPN, if applicable.**	
Q2	**Au: Please check the correction made to the sentence** 'These alleles...' and approve.	
Q3	**Au: Please provide forename/initial for** 'Olsen'.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by ⌐ or ⌐(x2)
Insert in text the matter indicated in the margin	⌐	
Delete	/ through single character, rule or underline or ⊢⊣ through all characters to be deleted	⌐/ or ⌐/(x2)
Substitute character or substitute part of one or more word(s)	/ through letter or ⊢⊣ through characters	new character / or new characters /
Change to italics	— under matter to be changed	⌐
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	~ under matter to be changed	~
Change to bold italic	≈ under matter to be changed	≈
Change to lower case	Encircle matter to be changed	≠
Change italic to upright type	(As above)	⌐
Change bold to non-bold type	(As above)	⌐
Insert 'superior' character	/ through character or ⌐ where required	Y or Y under character e.g. Y^{3} or Y^{3}
Insert 'inferior' character	(As above)	⌐ over character e.g. Y_{2}
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	Y or Y and/or Y or Y
Insert double quotation marks	(As above)	Y or Y and/or Y or Y
Insert hyphen	(As above)	⊢
Start new paragraph	⌐	⌐
No new paragraph	⌐	⌐
Transpose	⌐	⌐
Close up	linking ⁀ characters	⁀
Insert or substitute space between characters or words	/ through character or ⌐ where required	Y
Reduce space between characters or words	| between characters or words affected	⊤