Theor Appl Genet (2004) 109: 596–608
DOI 10.1007/s00122-004-1658-0

ORIGINAL PAPER

B. Barrett · A. Griffiths · M. Schreiber · N. Ellison · C. Mercer · J. Bouton · B. Ong · J. Forster · T. Sawbridge · G. Spangenberg · G. Bryan · D. Woodfield

A microsatellite map of white clover

Received: 17 September 2003 / Accepted: 8 March 2004 / Published online: 22 April 2004
© Springer-Verlag 2004

Abstract The white clover (*Trifolium repens*) nuclear genome (n=2x=16) is an important yet under-characterised genetic environment. We have developed simple sequence repeat (SSR) genetic markers for the white clover genome by mining an expressed sequence tag (EST) database and by isolation from enriched genomic libraries. A total of 2,086 EST-derived SSRs (EST-SSRs) were identified among 26,480 database accessions. Evaluation of 792 EST-SSR primer pairs resulted in 566 usable EST-SSRs. Of these, 335 polymorphic EST-SSRs, used in concert with 30 genomic SSRs, detected 493 loci in the white clover genome using 92 F_1 progeny from a pair cross between two highly heterozygous genotypes—364/7 and 6525/5. Map length, as estimated using the JOINMAP algorithm, was 1,144 cM and spanned all 16 homologues. The *R* (red leaf) locus was mapped to linkage group B1 and is tightly linked to the microsatellite locus *prs318c*. The eight homoeologous pairs of linkage groups within the white clover genome were identified using 96 homoeologous loci. Segregation distortion was detected in four areas (groups A1, D1, D2 and H2). Marker locus density varied among and within linkage groups. This is the first time EST-SSRs have been used to build a whole-genome functional map and to describe subgenome organisation in an allopolyploid species, and *T. repens* is the only Trifolieae species to date to be mapped exclusively with SSRs. This gene-based microsatellite map will enable the resolution of quantitative traits into Mendelian characters, the characterisation of syntenic relationships with other genomes and acceleration of white clover improvement programmes.

Communicated by C. Möllers

B. Barrett (✉) · A. Griffiths · N. Ellison · C. Mercer · G. Bryan · D. Woodfield
AgResearch Limited, Grasslands Research Centre,
Private Bag 11008 Palmerston North, New Zealand
e-mail: brent.barrett@agresearch.co.nz
Tel.: +64-6-3518094
Fax: +64-6-3518032

M. Schreiber
AgResearch Joint Bioinformatics, Institute School of Biological Sciences, University of Auckland,
3a Symonds St,
Private Bag 92-019 Auckland, New Zealand

J. Bouton
Department of Agronomy, University of Georgia,
Athens, GA 30602, USA

B. Ong · J. Forster · T. Sawbridge · G. Spangenberg
Department of Primary Industries Plant Biotechnology Centre,
La Trobe University,
Bundoora, Victoria, 3086, Australia

Introduction

White clover (*Trifolium repens*) is a source of quality animal fodder that is commonly found in pastoral agricultural systems in the temperate latitudes. This plant species, in symbiosis with bacteria of the genus *Rhizobia*, transforms atmospheric di-nitrogen into plant-available forms, thereby promoting more sustainable and productive agriculture. It is a member of the tribe Trifolieae, which contains, among others, the genera *Trifolium*, *Medicago* and *Melilotus* (Gillett 1985; Williams 1987). The basic chromosome number (x) of this tribe is eight, although in some species x=7, 6 or 5 (Williams 1987). Although allopolyploidy is infrequent in the Trifolieae, white clover exhibits allotetraploid (n=2x=16) transmission, with highly regular bivalent formation (Atwood and Hill 1940). The physical size of the genome is approximately 956 Mbp, about twice that of the model legume *Medicago truncatula* (n=x=8) at 466 Mbp (Bennett and Leitch 2003). The white clover karyotype consists of a homogeneous set of chromosomes, however 2 of the 16 are distinguishable by the presence of satellited knobs (Chen and Gibson 1970). While the evolutionary origin(s) and centre(s) of white clover diversity await conclusive demonstration, derivation from two progenitor diploid *Trifolium* species is likely (Badr et al. 2002).

The white clover mating system is regulated by a highly polymorphic, gametophytic self-incompatibility system

with outcrossing in ambient conditions approaching 100%, resulting in highly heterogeneous populations of highly heterozygous individuals. Despite the complication of mating system and polyploidy, studies of the *S* (self-incompatibility) (Atwood 1940), *Ac* and *Li* (cyanogenesis) (Atwood and Sullivan 1943), *R* (red leaf) (Corkill 1971) and *V* (leaf mark) (Carnahan et al. 1955) loci in white clover are benchmarks in single-locus genetic analysis of legumes.

The recent development of ubiquitous genetic marker technologies in conjunction with increased computational capability has provided abundant resources for whole-genome linkage analysis. Among the suite of DNA marker systems available, microsatellites or simple sequence repeats (SSRs) are a preferred technology for plant genome analysis (Morgante and Olivieri 1993). SSRs occur at a high frequency (Li et al. 2002) and exhibit high mutation rates (Vigouroux et al. 2002), preferential association with non-repetitive regions of the genome (Morgante et al. 2002), codominant inheritance, inter-specific transportability (Decroocq et al. 2002; Eujayl et al. 2004; Griffiths et al. 2002), a high-throughput assay speed and cost efficiency. Evidence is accumulating which suggests that most SSRs are not neutral genetic markers, do undergo selection and as such are not randomly distributed in the genome on the basis of size, motif, or polymorphism (Li et al. 2002). There is also substantial evidence to suggest that motif, length and mutation are dependent on whether the SSR resides in non-coding, coding untranslated or coding-translated DNA (Morgante et al. 2002). A further barrier to the use of SSRs is attrition associated with their development from genomic DNA libraries (Squirrell et al. 2003). Despite these concerns, SSRs have been used to map eukaryotic genomes (Sharopova et al. 2002), including allopolyploid plants (Roder et al. 1998). The attrition problem may be mitigated by in silico SSR discovery using expressed sequence tag (EST) or other sequence databases (Cho et al. 2000). EST-derived SSRs (EST-SSRs) occur at a high frequency in the genome (Kantety et al. 2002) and have greater inter-specific transferability than genomic SSRs, with the difference becoming more pronounced with increasing phylogenetic distance (Griffiths et al. 2002). EST-SSRs also exhibit high levels of polymorphism despite their association with sequences experiencing direct selection pressure (Cho et al. 2000; Eujayl et al. 2002).

The advantages of using gene-associated sequences as genetic markers include the ability to utilise bioinformatic resources and whole-genome sequencing initiatives for comparative genome analysis (Fulton et al. 2002; Gualtieri et al. 2002; Holton et al. 2002; Lan et al. 2000; Zhu et al. 2003), to construct function orientated maps (Matthews et al. 2001; Wu et al. 2002) and to identify gene candidates for the direct investigation of sequence function relationships (Tasma and Shoemaker 2003; Thornsberry et al. 2001). Despite being non-neutral genetic markers with non-random physical distribution (Li et al. 2002; Wu et al. 2002), gene-associated markers have been used to create comprehensive linkage maps in several species (Davis et al. 1999).

Several whole-genome linkage maps in Trifolieae taxa have been developed. Existing genetic maps of diploid and tetraploid *Medicago sativa* and the model species *M. truncatula* provide comprehensive marker resources for both species and have created a standardised nomenclature for genetic marker analysis within this tribe (Brouwer and Osborn 1999; Choi et al. 2004; Julier et al. 2003; Kalo et al. 2000; Thoquet et al. 2002; Zhu et al. 2002). Recently, a comprehensive genetic map in red clover (*T. pratense*, x=7) has been developed using restriction fragment length polymorphism (RFLP) in a backcross population (Isobe et al. 2003). An initial white clover genetic map utilising genomic SSRs to aid the placement of amplified fragment length polymorphisms (AFLPs) in an F_2 population derived from a pair cross between partial inbreds has also recently been developed (Jones et al. 2003). Although Jones et al. (2003) has provided some insight into the white clover genome, this low-density map (mean 6.1 cM/marker) was not able to describe the entire genome and could not ascertain the homoeologous relationship between the subgenomes.

We report the in silico discovery of white clover EST-SSRs and their use in combination with genomic SSRs to construct a comprehensive genetic linkage map of the allotetraploid white clover genome. This is the first report of EST-SSRs being used to build a whole-genome functional map and to describe subgenome organisation in an allopolyploid species and the only example found in the Trifolieae of an exclusively SSR-based genetic map.

Materials and methods

Plant material

A double-pseudo testcross population of 92 F_1 progeny was created by a manual pair cross between two phenotypically divergent, highly heterozygous white clover (*Trifolium repens*) genotypes—6525/5, a parent of the cultivar Grasslands Sustain (Caradus et al. 1997), and 364/7, an elite plant selected from the nematode resistance recurrent selection programme at AgResearch (Mercer et al. 1999). Parent 6525/5 exhibits small to medium leaves, many long thin stolons and dense growth, whereas 364/7 exhibits medium leaves, few short thick stolons and open growth. Parent 6525/5 is heterozygous at the *R* locus ($R^f r$), whereas 364/7 is homozygous (*rr*); both lines are cyanogenic. DNA was purified from immature leaf tissue using the DNeasy kit (Qiagen, Valencia, Calif.).

Genomic SSR discovery

SSRs identified from DNA libraries enriched for di-, tri- and tetranucleotide motifs were utilised. White clover SSR-enriched libraries were constructed in the plasmid vector pUC19 by Genetic Identification Services (Chatsworth, Calif.). Individual colonies were transferred into 50 µl Luria-Bertoni medium, and 1-µl aliquots were used for PCR analysis. Depending on the level of enrichment for particular SSR motifs, clones were either picked at random for PCR using both flanking M13 forward and reverse vector primers or pre-screened by PCR using individual flanking vector primers in combination with a repeat primer specific for the SSR repeat type of

interest. The PCR was performed in a 20-µl volume containing 1× AmpliTaq buffer (Applied Biosystems, Foster City, Calif.), 0.2 m*M* of each dNTP, 0.2 µ*M* of each primer, and 1.25 U of AmpliTaq DNA polymerase (Applied Biosystems). The thermocycle parameters consisted of an initial denaturation at 94°C for 4 min followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 60 s, with a final extension of 72°C for 7 min. The PCR products were electrophoresed on a 1.2% SeaPlaque (Cambrex, Rockland, Me.) agarose gel in 1× TTE buffer (89 m*M* Tris-HCl, 29 m*M* taurine, 0.54 m*M* EDTA) and excised from the gel under long-wavelength UV. DNA was purified from the excised fragment by heating at 65°C for 5 min then snap-freezing in liquid N_2, followed by centrifugation for 10 min at room temperature. A 4-µl aliquot of supernatant was sequenced using the M13 forward or reverse primer and the BigDye terminator (Applied Biosystems) cycle-sequencing kit. Sequencing reactions were resolved on an ABI 3700 capillary electrophoresis array (Applied Biosystems). Primers were designed to the SSR flanking regions using PRIMER DESIGNER 4 software (Scientific & Educational Software, Durham, N.C.).

EST-SSR discovery

Accessions in a white clover EST database (Sawbridge et al. 2003) were mined using a recursive algorithm implemented in Sputnik (http://rast.abajian.com/sputnik/). Di-, tri-, tetra- and pentanucleotide SSR motifs exhibiting three or more tandem repeat units and a minimum length of 11 nucleotides were identified. For each SSR with an adequate flanking sequence, forward and reverse primers were designed to a T_m of 60°C using PRIMER1 (http://wwwgenome.wi.mit.edu/ftp/distribution/software/). ESTs with SSRs were compared to TIGR *Medicago truncatula* (*Mt*)GI 7.0 database and GenBank 135.0 using BLASTN; query parameters were -e 0.000001 -v 10 -b 10.

SSR genotyping

SSR forward primers were modified by 5′ concatenation of the 18-mer 5′-TGTAAAACGACGGCCAGT-3′, which permitted concurrent fluorescence labelling of PCR products by a third primer with an incorporated fluorophore (Schuelke 2000). The 7-mer 5′-GTTTCTT-3′ was appended to the 5′ end of each SSR reverse primer to promote non-templated adenylation of amplicons and thereby simplify automated data acquisition (Brownstein et al.

Table 1 Thirty white clover SSR markers derived from CA_n-, ATG_n- and $CTAT_n$-enriched genomic DNA libraries. These markers identified 48 genetic loci in the F_1 white clover mapping population 346/7×6525/5

SSR ID	SSR motif	Calculated size (bp)	Linkage group	Forward primer (5′→3′)	Reverse primer (5′→3′)
ats002	ATG	253	A1	ggcgagagtgagacagtgga	tgtcgctccaaccaccagta
ats003	ATG	181	D1	cctcaagtccaccacctgtc	accacaccattcttgcttct
ats005	ATG	297	F1	tgaggaagacacgcctatca	gaatccaggagttatatcag
ats006	ATG	325	C2	tccatagccgtccgatgatt	actgaacgagccattgattg
ats029	ATG	286	A1, A2	gctccttccattacggtgtt	atcgctctcgctctccttct
ats032	ATG	196	A1, A2	gggcagccattgtcattc	tcgcagttattccgatcgatt
ats041	CA	359	B1	tgttgtcggtcgtcttcact	atgtcgttgaaccgctagtg
ats054	CA	278	A1	gacaccgattatgtgcaaga	aatcacgacgagcgacaaca
ats055	CA	392	B2, E2	caatacaatcaccgcaccag	tctctgcttcgcgtcttctc
ats058	CA	267	E1	caatcaactctgctaagtgt	aaggagagtatgtgaagtag
ats066	CA	236	H1, H2	tattcaccacacgcctctac	atgaggagaaggcaggagat
ats067	CA	405	B2	ttactgatgcggttagtgtt	cctccttgctgttgttatcc
ats070	CA	278	C1, C2, G2	gtcattggtgatggtgttct	tttcgtcagtggcggtgctc
ats072	CA	270	B1, D1	cttcccatatgtggatgaatag	tggcacagcagaagcaataa
ats073	CA	302	F1	gagcaactggcaatacataa	ttgttggaactgctgataga
ats075	CA	205	B1	actcgatcaccatgtgagtc	tcataaccggcgtggaagaa
ats084	CA	275	B1	aaccaagcgtcactacttca	ggtgttccgattctattctg
ats099	CA	301	C1	aaccaagcgtcactacttca	atgcggatggttgttgttcg
ats113	ATG	225	D2	tgtgagctggtgaattgagt	ggaggtgatgatctctatcc
ats121	ATG	248	B2	tccaccgctgttgtgcaacc	ggatcgcattctccatctcc
ats123	CTAT	450	F1, F2	acacaattcagcagagattc	cagtcgttggatcgtagtaa
ats125	CTAT	276	E1	atctaaggctccaaagtatc	gcaacaatagaagcagcaatca
ats126	CTAT	232	F1	acacaattcagcagagattc	cctggttactgataagttac
ats131	CA	302	B1, B2	atgatcctccggaccggatg	cgcacggtggacgttgtaag
ats153	CA	296	B1	taatgtccaagtggttagag	ttcacttgctgctgctggta
ats176	CA	300	F1, F2	atcagttggcggttcagtag	gtcaatgtggcagctcttgt
ats186	CA	154	H1, H2	tgatggcgcaatcaggaatg	aacgcacacagccctagttt
ats205	CA	226	B1	acaccgtcgttatgaagtat	gacttgtgactactgataga
ats226	ATG	313	B1	catctactcaccaccaccta	cagcagcagcagcagcgata
ats227	ATG	211	D1	tctgaatcagtcggtgtagc	gaggaattgccacgatgatg

1996). The PCR protocol of Schuelke (2000) was modified with a 30 min final extension to promote non-templated adenylation of the amplicons. PCR products were resolved in an ABI 3100 capillary electrophoresis array (Applied Biosystems). Electropherograms were analysed using ABI Prism GENESCAN 3.7 and GENOTYPER 1.7 software (Applied Biosystems).

Genetic linkage analysis

The analysis of maps in each parent and the subsequent development of a consensus map of the genome were carried out using JOINMAP ver. 3.0 (http://www.kyazma.nl). Initially, segregating peaks were scored into a 0/1 data matrix and subsequently recoded as single-marker loci (*ao* × *oo* for 364/7 and *oo* × *ao* for 6525/5). For single-parent analysis, loci grouping and ordering probability minima were LOD=5 and LOD=2, respectively.

The two single-parent maps were fused to create a bi-parental consensus map of the white clover genome. Segregation data from markers polymorphic in both parents with two or three segregating alleles per locus (segregation type *ab* × *ab* and *ab* × *ac*, respectively) were recoded prior to initial linkage analysis and incorporated into the dataset as bridging loci to identify the homologous pair for each consensus linkage group. Locus pairs exhibiting two alleles in each parent, detected by the same SSR marker, and ordered in similar positions on matching parental homologues were paired and recoded as type *ab* × *cd* prior to estimation of the consensus map. Prior to fusion, homologues from each parent were screened for similarity of recombination frequency (map length) and surveyed for differential segregation distortion. For 15 of the 16 linkage groups, a consensus map combining parental homologues was estimated at grouping and ordering probability minima of LOD=8 and LOD=2, respectively, and genetic distances were estimated using the Kosambi mapping function. For parental groups G2 no bridging loci are available, so group identity was confirmed on the basis of loci on G2 that were homoeologous with loci on consensus group G1. The eight pairs of homoeologous linkage groups were identified and aligned using putative homoeologous loci, which were declared when an SSR detected loci at similar map positions in two homologues. The most robust pairings of homoeologous loci were used to designate homoeologous pairs of linkage groups. The eight homoeologous pairs were arbitrarily named A–H, and homoeologues within each pair were arbitrarily named one and two, respectively.

Table 2 Thirty-two white clover SSR markers derived from an EST database. Two markers per linkage group were selected based on multiple criteria including genome position, peak quality and probability of being a single-locus marker

SSR ID	SSR motif	Calculated size (bp)	Linkage group	Forward primer (5→3′)	Reverse primer (5→3′)
prs006	CTT	247	G2	cctggaactctgaactcgtacc	agtgaatggagaagaagtgtgatg
prs055	GAA	155	C1	taatagcggccatatcggag	aaaaaaaggcaacacgcagt
prs100	GTT	227	A2	caatcttaaatggtggaggagc	ctacaccacaaacaacattgttagc
prs129	CCA	152	H2	ccgtgattgcatatccagtg	cgccacccttgttagttgtt
prs156	TTCT	190	B2	gcatcagtgtcagaagccaa	caaagagaaaaggtggttttgg
prs186	AATTC	251	A1	gtgagcgagcactgtttctg	gccatgcacttgttttgaga
prs203	TGTT	215	G1	aacgatccgatcttgattgc	actcatttttccctgcatgg
prs247	AAG	259	D2	tcatcttcatcaacagtttccg	cttcccttctatctctcatgttaacc
prs251	AGT	225	H2	tggagaaattggtggtgtca	ttccacaacaccatctcatca
prs256	TCTG	259	E1	ccgttttcgttctcgaagag	gagtggagaggaagtcgtcg
prs264	ACC	215	C1	tcatcttcttcaccagcacg	cggtgaaactgttcctggtt
prs268	GAAT	172	F1	gctgagaaagcaagattaggttg	ccagaaacaaagaacagcagc
prs279	ATG	164	H1	atggctaatgcaaaattggc	tagagactgcagatcatcatctcc
prs285	AAC	248	E2	tcctcaaacgacctcgttct	tccttccaggtcttgcagtt
prs305	AAC	153	G2	tcaaactcgacgaagaagacg	gatggagatggagattcgga
prs344	TA	267	F1	tcaaagaaagcaaaggaagagc	ttccattcttcacgactcttca
prs367	TGAAC	236	C2	agattctctgcatctttttcgc	aagctccaaaccctatcggt
prs369	CCACT	252	B1	atcatcccaatcatcttcatcc	cctaatcttcgtagatcgtcgg
prs408	TCT	171	E1	actaagaaggatctgaatctctctgc	tcactgatccagtggtagatgg
prs426	TTG	194	A2	ttattgatatcggaagcgacg	gttctcatgcgacgctacaa
prs427	CTT	196	D1	gttcttcaatctccacaatacgc	taaaggaaaaggtgaggatagtgg
prs433	TTC	202	C2	cccaaccgcataacacttct	aaagatgaggagaagatcaaggg
prs461	TTC	239	A1	accttccgatatcccaaacc	atggtgcgtttggagatagg
prs499	CCT	138	E2	tatgactagccccgtcatcc	aattccttcggggaagttgt
prs510	TTC	186	G1	gctgctctcccttttttcct	ggaacatttgcagggaggta
prs582	GAA	267	B1	ccggttcgattcaacaagtt	ctgcagatccagtaatgatttcc
prs599	TAA	207	D1	ttgaagttccacaaagatgtgc	aaacgacgtcaccaaccttc
prs612	ATG	174	D2	ttgaactagtcgttggatggg	gagagggtttcaggaacatacg
prs645	TAT	159	F2	acccattttcatttcccaca	ccattcttccctgagctttg
prs651	GAA	160	H1	gtaacgattctgttgtttacgtcg	ctcgtcggtaaacgaatactcc
prs653	ACC	232	F2	ggagcagcaccagaagaaac	aggttgaagtggtggaggtg
prs734	TCT	202	B2	aactcttcgacaatggcttca	tgaactgttgcgcctagatg

Results

SSR discovery

A total of 107 white clover genomic SSRs were identified in enriched DNA libraries, including 68 $(CA)_n$, 22 $(ATG)_n$, and 17 $(CTAT)_n$ repeats. Of these 107, 30 (28%) provided robust PCR products and detected polymorphism between or within the parent genotypes (Table 1). Trimer (ATG) genomic SSRs were the class most likely to be informative, with 10 of 22 (45%) exhibiting polymorphism; tetramer (CTAT) repeats provided the least informative markers, with 18% of the total detecting polymorphism in this mapping population.

White clover SSR markers were also developed from an EST database (Sawbridge et al. 2003). A scan of 26,480 sequence accessions in the EST database revealed 2,294 EST-SSRs in silico within 1,841 accessions (mean=1.25 SSRs/positive accession; range: 1–7 SSRs/positive accession). Primers were designed for 2,086 EST-SSRs in 1,697 accessions, including 264 di-, 1,353 tri-, 263 tetra- and 206 pentanucleotide motifs. EST-SSRs ranged in length from 11 to 76 nucleotides (mean=16), and the number of SSR units ranged from 3 to 25, with a median of four (n=700) (Fig. 1). When compared to the *Mt*GI 7.0 database (http://www.tigr.org), 1,343 (79%) accessions for which a primer pair had been designed exhibited homology (1.0E-7) to an annotated sequence. Among the *Trifolium* ESTs with no homology to *Medicago truncatula* ESTs, 51 (3%) had significant hits in *Arabidopsis thaliana* (*At*GI 10.0) or GenBank (release 135.0) databases. The remainder may represent either undiscovered transcripts in other EST projects or sequences unique to the white clover genome. Primer pairs for 792 EST-SSRs were synthesised and evaluated, including 116 di-, 538 tri-, 69 tetra- and 69 pentanucleotide repeats. Although database annotations and motif type were known for the EST-SSRs, repeat length was the only selection criterion used in order to minimise genome coverage bias. Primer pairs were screened for amplification using genomic DNA from the parental genotypes 364/7 and 6525/5. A total of 566 (71%) of the EST-SSR primer pairs exhibited robust amplicons when assayed at design PCR parameters. A subset of 32 EST-SSRs is described in Table 2. Neither SSR length nor motif type was indicative of the likelihood of a particular EST-SSR to produce an amplicon. Realised amplicon size was usually close (mode=99%) to the size predicted in

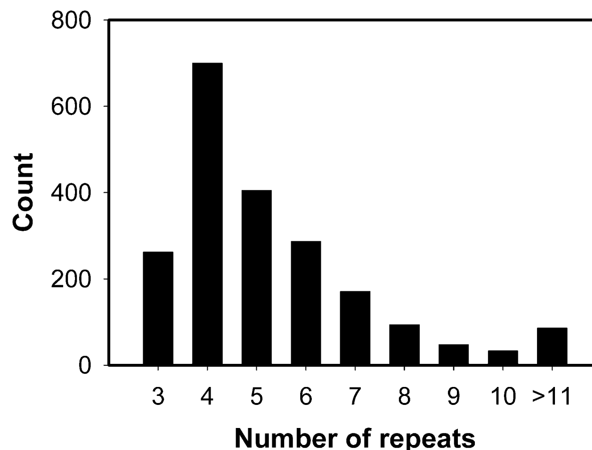

Fig. 1 Distribution of 2,086 white clover EST-SSRs on the basis of number of repeats units detected in silico. The numbers range from 3 to 25 repeat units

Table 3 Simple sequence repeat loci distribution among the 16 linkage groups of the white clover genome. Linkage data were developed using 92 F_1 progeny in the mapping population 364/7×6525/5. Locus segregation distortion was declared at P<0.05

Linkage group	Length (cM)	Number of SSR loci	Number of loci polymorphic			Number of loci distorted
			364/7	6525/5	Both	
A1	76	35	11	12	12	3
A2	77	38	10	16	12	2
B1	70	41	15	15	11	0
B2	69	39	11	17	11	0
C1	69	38	8	18	12	1
C2	62	20	6	5	9	0
D1	82	52	22	12	18	9
D2	94	47	23	12	12	5
E1	91	31	8	13	10	3
E2	72	27	5	16	6	2
F1	72	22	8	8	6	0
F2	76	18	5	10	3	0
G1	77	26	5	14	7	1
G2	28	14	11	3	0	1
H1	69	22	10	7	5	1
H2	60	23	14	6	3	6
Total	1144	493	172	184	137	34
Mean	71.5	31.1	10.8	8.6	8.6	2.1

silico for the 566 usable primer pairs, and 80% (430) of the EST-SSRs exhibited amplicon mean sizes ±20% of the predicted size. However, the overall mean amplicon size detected was 128% of the predicted size. This bias towards longer-than-predicted products may be attributed in some cases to the presence of introns in the genomic DNA template. Subsequent sequencing of two longer-than-predicted EST-SSR amplicons revealed putative introns (data not shown).

Among the 566 usable EST-SSRs, 379 (67%) detected polymorphism between or within the parental genotypes. Linear regression analysis indicated that the number of SSR units in an EST was not correlated with the likelihood of an SSR being polymorphic within or between these two genotypes (p=0.448 for EST-SSRs of 3–10 repeat units). Sample sizes for EST-SSRs with 11–25 repeats were too small to implement regression analysis. Motif type was also compared with the likelihood of an EST-SSR detecting polymorphism, and these likelihood values ranged from 61% of the tetramers to 70% of the dimers.

Genetic linkage analysis

We used 335 EST-SSRs and 30 genomic SSRs to genotype the two parents and 92 F_1 progeny of the white clover population 364/7×6525/5 and detected 493 SSR loci. Of these 493 loci, 172 were polymorphic in 364/7 only, 184 were polymorphic in 6525/5 only and the remaining 137 (28%) were polymorphic in both parental genotypes (Table 3). Among the loci polymorphic in both parents, 20, 40 and 77 loci exhibited two (*ab* × *ab*), three (*ab* × *ac*) and four (*ab* × *cd*) alleles, respectively. Forty-four loci were detected by the 30 genomic SSRs (mean=1.47 loci/SSR), which is higher than the mean of 1.34 loci/SSR for EST-SSRs. A higher proportion of genomic SSRs than EST-SSRs exhibited three or four different alleles per locus (35% vs. 22%).

Linkage analysis was initiated in each parent using single-dose marker data. This independent analysis of the 309 loci in 364/7 and the 323 loci in 6525/5 resulted in assembly of 16 linkage groups in each parent, ranging in length from 7 cM to 94 cM and containing 3–35 loci (data not shown). These single-parent maps were subsequently assembled into a bi-parental consensus linkage map of the white clover genome consisting of 16 linkage groups. For all homologues except G2, one or more of the 40 loci with three alleles (*ab* × *ac*) were used to unambiguously identify homologous pairs between the parental maps. G2 parental linkage groups had no loci that were informative in both parents; however, both had homoeologous loci informative in G1 that were used to confirm the identity of group G2 in each parental map build. Nine EST-SSR loci segregating from 364/7 had homoeologous loci on the lower portion of group G1, whereas two of the three loci segregating from 6525/5 group G2 had homoeologous loci on the upper segment of G1. The 77 loci pairs exhibiting two alleles in each parent, detected by the same SSR marker, and ordered in similar positions on matching parental homologues were paired and recoded as type *ab* × *cd*. Finally, the 20 loci with two allele segregation types (*ab* × *ab*) were ordered onto the consensus linkage groups. In regions of differential segregation distortion between the parent-specific maps, loci were validated in each parental map and the bi-parental consensus map to ensure that no spurious re-ordering of markers was caused in the consensus integration. In only one homologous pair (D2) did segregation distortion, originating from 6525/5, cause spurious ordering in the consensus estimate; thus, the 12 loci from 6525/5 were treated as grouped accessory markers and were not ordered in the final map build.

The final genome map estimate included 493 SSR loci and one morphological marker distributed across 16 linkage groups and spanning 1,144 cM of the white clover nuclear genome (Fig. 2). Five loci remained unlinked, three originating from 364/7 and two from 6525/5. Linkage groups ranged from 60 cM to 94 cM in length and contained 14–52 loci per group (Table 3). Genetic map density ranged from 1.6 cM to 4.2 cM per locus among the 16 groups (overall mean 2.3 cM per marker), indicating that this is a medium-density map. The 16 genetic linkage groups presumably correspond to the 16 bivalents observed in the white clover mitotic karyotype.

Thirty-two loci exhibited significant (p<0.05) segregation distortion. Eleven distorted loci were detected in 364/7 only, 14 in 6525/5 only and nine in both parents. Homoeologous pair D accounted for 14 distorted loci, primarily from 6525/5, while groups A1 and H2 exhibited distinct clusters of distortion at the top of each group, with distorted markers originating from both parents. The remainder of the distorted markers were scattered throughout the genome.

Ninety-six SSRs detected homoeologous loci between homoeologous linkage groups, ranging from four (F) to 14 (B) loci per pair of linkage groups, thereby facilitating unambiguous identification of the eight homoeologous pairs in the allotetraploid white clover genome. Twenty-seven percent of the EST-SSRs detected homoeologous loci, whereas only 20% of the genomic SSRs did. Marker order and inter-locus gap sizes were well conserved between homoeologues, with only minor discrepancies. A notable exception is pair G, which showed a substantially different marker ordering and inter-locus gaps between G1 and G2, suggesting that additional marker loci on group G2 from both parents are necessary for clarification of genetic order. Twenty-five (6.8%) SSRs detected duplicate non-homoeologous loci, both within and between linkage groups (e.g. *prs462* detected two loci on B2; *ats055* detected two loci, one on B2 and one on E1).

Loci were not evenly distributed among homologues and homoeologous pairs, ranging from 18 (F2) to 52 (D1) per homologue and 40 (F, G) to 99 (D) per homoeologous pair (Table 3). Variable marker-locus densities also were evident within homoeologous pairs (e.g. C1 vs. C2) (Fig. 3). The consensus map has 464 inter-locus intervals (mean=2.5 cM, median=1.3 cM) with a maximum of 16.4 cM on group C2 (Fig. 4). The maximum gap figure

0 10 20 30 40 50 60 70 80 90 100

A

1: prs051d, ats002x, prs079x, prs380x, prs398c, prs365y, ats054x, prs200x, prs687c, prs618b, prs561a, prs021z*, prs713a, prs190x*, prs461a, prs786c, prs526b, prs306z*, prs770b, prs473a, prs045b, prs146b, prs377b, prs702x, prs717b, prs497d, prs738a, prs002c, prs396d, prs010a, prs643x, ats029d, ats032x, prs089b, prs186b (76)

2: prs079x2, prs467a, prs380c, prs051a, prs365b, prs200x2, prs202b, prs426c, prs673d, prs190x, prs595x, prs751a, prs561x, prs725a, prs364b, prs306b, prs786a, prs546a*, prs770x, prs497b, prs046b, prs504b*, prs240c, prs521a, prs112x, prs550c, prs010x, ats032x2, prs643e, prs276b, ats029a, prs685x, prs530c, prs100a, prs363c, prs357a, prs089x, prs753z (77)

B

1: prs312a, prs544a, prs582b, prs579x, prs431x, prs139z, prs590x, prs701a, prs349x, prs289b, ats075a, prs313c, prs597c, ats226x, prs531d, prs149z, prs335d, ats041a, prs361a, prs456a, prs217e, prs005x, prs402c, prs318c, R, prs468a, ats153a, prs659x, prs119b, prs369a, prs217a, ats084b, prs007x, prs014x, ats131a, ats205b, prs485b, prs562f, prs371d, ats072c, prs616a, prs568a (70)

2: prs431a, prs077a, prs001x, ats121b, prs544c, ats055a, prs646b, prs683c, prs044x, prs139b, prs701d, prs531c, ats067a, prs603b, prs309b, prs456c, prs361x, prs156a, prs462z, prs462b, prs418c, prs153b, prs659a, prs119e, prs749x, prs600x, prs675c, prs146d, prs007x2, prs014x2, ats131d, prs562x, prs687d, prs371b, prs568d, prs734z, prs748c, prs381x, prs715b (69)

C

1: prs609a, prs615x, prs720b, ats070x, prs588d, prs055b, prs671c, prs122a, prs670e, prs722z, prs383b, prs784d, prs538z, prs639x, ats099b, prs269a, prs037a, prs604x, prs565b, prs637x, prs216c, prs435b, prs490a, prs473d, prs107x, prs379a, prs414b, prs705d, prs516a, prs394a, prs737x, prs264b, prs115c, prs730x, prs158x, prs315b, prs294x, prs060f (69)

2: prs615x2, prs670x, ats070x2, prs588b, prs722x, prs367z, prs784c, prs783c, prs639c, ats006x, ats070j, prs433a, prs386x, prs435a, prs216a, prs161x, prs326b, prs543x, prs060g, prs737e (62)

D

1: prs513a, prs421b, prs613d, prs108b, prs178a, prs360a, prs255a, prs308x, prs668b, prs592x, prs599b*, prs648a*, prs207b*, prs379g, prs411x, ats003a*, prs282a, prs267z, prs226x, prs073x, prs024a, prs008a*, prs521e, prs322b, prs9326c, prs134a, prs573d, prs340c, prs300a, prs166a, prs017f, ats072e, ats227x, prs180x, prs494d, prs640x, prs191c*, prs598x, prs271a, prs385x*, prs681b*, prs427a, prs334z1*, prs330x, prs711x, prs520x, prs350f, prs583z, prs387x, prs124x, prs122d, prs373c (82)

2: prs308z, prs513g, prs613a, prs165b, prs108x*, prs612c, prs283x, prs592e, prs666b, prs379c, prs320b, prs282b, prs525a, prs606x*, prs790a, prs029c, prs640a, prs166b, prs180e, prs494b, prs160b, prs017a, prs334z2*, prs523a, prs181b, prs334f, ats113x, prs330x2, prs564d, prs247a, prs350x, prs509b, prs090x, prs124x2, prs373x (95)

E

1: prs035b, prs663a, prs236a, ats058x, prs331x, prs408a, prs284b*, prs225a, prs012x, prs757z, prs690z2, prs311c, prs150b, prs484c, prs085x, ats055b, prs406x, prs384x, prs783b*, prs634d, prs368c, prs460a, prs035x, prs642d*, prs204f, prs256a, ats125b, prs762x, prs391c, prs271b, prs556b (91)

2: prs664c, prs704c, prs641a, prs331a, prs285a, prs632c, prs225g, prs284e, prs498a, prs144b, prs405b, prs234x, prs690z1, prs552b, prs214a, prs085x2, prs484a, prs619a, prs499a, prs762c, prs511x, prs177z*, prs406g, prs754b, prs709a, prs634a, prs528x* (72)

F

1: prs500b, prs021d, prs268b, prs343a, ats005x, ats176x, prs682b, prs321x, ats123b, ats126e, prs353a, prs419b, prs157b, prs596x, prs710a, prs344a, prs566x, prs700b, prs465c, ats073x, prs628b, ats126b (72)

2: prs169y, prs708a, prs645d, prs490c, prs328c, ats176d, prs328a, prs184b, prs474x, prs145a, prs764b, prs131b, ats123c, prs353y, prs653b, prs731b, prs465e, ats126c (76)

G

1: prs204x, prs209f, prs781d, prs072x, prs575c, ats070f, prs763d, prs442c, prs443x, prs310x, prs120b, prs299b, prs203a, prs608z, prs022c, prs554c, prs075x, prs487b, prs022e, prs779a, prs510d, prs662d, prs064b, prs479x, prs506c, prs479e* (77)

2: prs006b, prs209g, prs575a, prs302a, prs443f, prs506b, prs305b, prs554a, prs608d, prs022a, prs479a, prs662b, prs726b, prs041a* (56)

H

1: prs545b, prs332b, prs336x, prs188c, prs279b, prs607x, prs157c, prs136b, prs407f, prs197b, prs342c, prs295d, prs154c, prs651b, prs647x, prs732b, prs534f, ats186x, ats066b, prs691a, prs329x*, prs221b (66)

2: prs332c*, prs336b*, prs545x*, prs407c, prs188e*, prs136a*, prs480a*, prs197d, prs295a, prs251b, prs342a, prs732a, prs129b, prs682a, prs712a, prs534g, prs290c, ats186x2, ats066a, prs138a, prs329x2, prs235c, prs221d (60)

◀**Fig. 2** A genetic linkage map of the white clover genome (n=2x=16), containing 493 SSR loci. The morphological locus "R" is mapped to group B1. The eight homoeologous pairs of linkage groups are arbitrarily labelled A–H, and homoeologues within each pair are designated one and two. Genetic length (cM) of each homologue is indicated *in brackets below* each group. Loci prefixes *ats* and *prs* denote genomic- and EST-SSRs, respectively. Loci detected by EST-SSRs with significant BLAST hits in the *Mt*GI7.0 database are *underlined*, whereas homoeologous loci are noted in *bold* and by *connecting lines* between the two homoeologues

excludes data from G2, which did not coalesce to form a bi-parental consensus and thus cannot be reasonably compared. SSRs also appeared to cluster within linkage groups (e.g. top and bottom of pairs B and C, top of pairs A and H). Genomic SSR markers were detected on all linkage groups except E2 and G2, with a bias towards homoeologous pairs B (15% of pair B loci were genomic SSRs) and F (23%) and away from D (4%). Within groups containing both SSR types, loci were intermingled (e.g. B1).

The adaxial red fleck, a cold-induced, leaf-epidermis-specific anthocyanin accumulation phenotype conditioned by the dominant R^f allele at the R locus exhibited a 1:1 segregation ratio among the progeny (χ^2=1.6, P=0.211) and was used as a morphological marker locus. Among the 92 F_1 progeny assayed, we did not observe any recombination between the R^f and *prs318c*, indicating that the two loci are tightly linked and that the observed alleles are in coupling phase on linkage group B1 in parent 6525/5. During evaluation of an additional 89 progeny we observed two recombinants between *prs318c* and R^f, providing an estimate of 1.1±0.006 cM between these two loci.

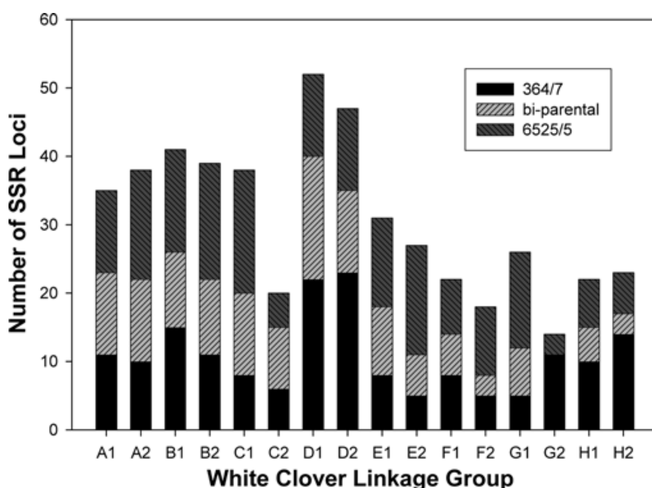

Fig. 3 A plot of the distribution of SSR loci among the 16 linkage groups of the white clover genome, with subdivision into those loci informative in genotypes 364/7, 6525/5 or both parental genotypes. Note that no SSR loci on group G2 are informative in both parents, which prevents the fusion of group G2 into a bi-parental consensus

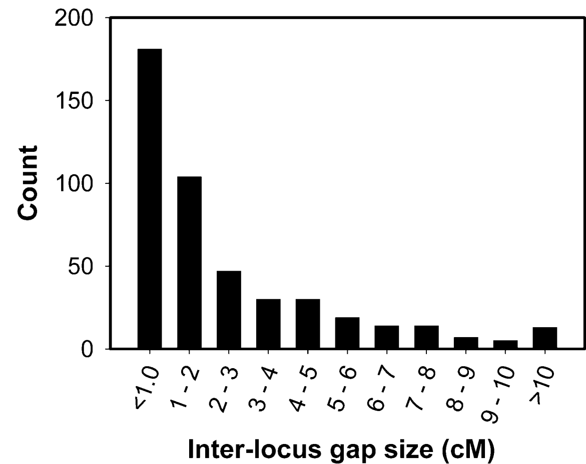

Fig. 4 Frequency distribution of inter-locus gaps (n=464) in a white clover genetic linkage map developed using SSR markers sourced from an EST database and enriched genomic libraries. Mean=2.3 cM per gap, mode=1.3 cM. The maximum gap is 16.4 cM on group C2. The gap between the two segments of G2 was excluded from analysis

Discussion

We developed microsatellite (SSR) molecular markers from enriched genomic DNA libraries and a white clover EST database. A subset of these were utilised to create a comprehensive map of the allotetraploid white clover genome by linkage analysis in the double-pseudo testcross population 364/7×6525/5.

Our EST-SSR initiative resulted in the discovery of one or more EST-SSRs in 7.0% of the database accessions. This density compares favourably with EST-SSR discovery programmes in *Vitis* (2.5%) (Scott et al. 2000) and a range of cereals (1.5–4.7%) (Kantety et al. 2002) but is lower than the 10.5% reported in *Capsicum* (Sanwen et al. 2000) and the approximately 20% reported in maize (Sharopova et al. 2002). Collectively, these data support the efficacy of SSR discovery in sequence databases and concur with the results of other investigations (Morgante et al. 2002) that SSRs occur at a high frequency in expressed regions of the plant genome. Among the five motifs we mined, trinucleotides formed the majority (65%) of the EST-SSRs identified. This is in close agreement with other observations in monocot and dicot plant species (Kantety et al. 2002; Morgante et al. 2002) and reflects the strong selection pressure against frameshift mutation in coding regions limiting the formation or mutation of non-trimer and non-hexamer SSRs (Metzgar et al. 2000). In contrast, non-coding regions of eukaryotic genomes have been shown to contain primarily dinucleotide repeats (Li et al. 2002).

Of the EST-SSRs that we evaluated, 71% produced PCR products after a single round of evaluation, a rate almost identical to the proportion obtained in a large-scale study of rice EST-SSRs (73%) (Temnykh et al. 2001) and lower than the 82% reported in *Lycopersicon* (He et al. 2003) and the 86% reported for genomic SSRs developed from enriched DNA libraries in white clover (Kolliker et

al. 2001). It does, however, compare very favourably with the low yield of our genomic SSR development effort (28%), which may be attributed to a high level of clone redundancy in the enriched DNA libraries (data not shown).

The relatively low attrition rate of SSR development in silico, when considered in conjunction with the ease of database mining versus library development (Squirrell et al. 2003), association of EST-SSRs with low-copy number sequences (Morgante et al. 2002) and high transportability between taxa (Eujayl et al 2004; Griffiths et al. 2002) provides further evidence for the benefits of in silico marker discovery, despite the slightly lower polymorphism rates. Although our EST-SSR utilisation rate may be increased by primer redesign, optimisation of the PCR or the use of a transcriptional representation of the target genome as a template for the PCR, it is more economic to continue development of new EST-SSRs from among the 1,294 that remain in silico. Once sequencing of the *Medicago truncatula* nuclear genome is complete and macrosyntenic relations between *Trifolium* and *Medicago* are described, these data may be used as a template for targeted development of the remaining EST-SSRs by implementation of a predictive bioinformatic approach such as COMPASS (Band et al. 2000) to identify ESTs with a threshold probability of mapping to under-represented or targeted regions in the white clover genome.

On average, the white clover EST-SSRs detected less polymorphism (fewer loci per SSR and fewer alleles per locus) than the white clover genomic SSRs. No other comparisons of EST-SSRs with genomic SSRs have been reported in a double-pseudo testcross architecture; however, diversity analyses have demonstrated that EST-SSRs were substantially less polymorphic than genomic SSRs among diverse germplasm, with the former in comparison with the latter showing a reduction of 50% in allotetraploid wheat (Eujayl et al. 2002) and 18% in tomato (Areshchenkova and Ganal 2002). Mutation rate and polymorphism levels are often correlated with SSR length (Sharopova et al. 2002; Temnykh et al. 2001; Vigouroux et al. 2002); however, the polymorphism potential of the 566 white clover EST-SSRs we assayed was not correlated with SSR length or repeat number (p=0.448). These data concur with observations suggesting that a threshold below the commonly used 20 nucleotide minimum may be appropriate for SSR discovery (Sanwen et al. 2000).

The polymorphism rate of 67% for white clover EST-SSRs in the 364/7×6525/5 mapping population is higher than that obtained in other EST-SSR studies (Holton et al. 2002) and higher than the 48% of genomic SSRs previously described in a F_2 white clover mapping population (Jones et al. 2003). The increase relative to cereal EST-SSRs may be attributed to white clover's promiscuous mating system, which is expected to drive polymorphism higher than a comparable inbred system (Awadalla and Ritland 1997; Dvorak et al. 1997). The increase relative to the genomic SSRs in the F_2 white clover mapping population may be attributed to the double-pseudo testcross population architecture, which samples up to four alleles per locus, and selection of genetically and phenotypically diverse parent genotypes. When considered together with other benefits (see below), the use of double-pseudo testcross structures in species with open-pollinated reproduction is an attractive strategy for whole-genome mapping.

Of the mapped white clover SSRs, 19% identified homoeologous loci, with a range of 5 to 16 per homoeologous pair. Applying the formula $(1/g)^{(n-1)}$ where g=numbers of linkage groups and n=numbers of syntenic loci (Reinisch et al. 1994), the probability of incorrect pair assignment ranges from 1.5E-5 to 8.7E-19 (pairs F1/F2 and D1/D2, respectively). The rate of homoeologous loci detection in white clover is similar to the rate of homoeologous loci detected in allohexaploid wheat using genomic SSRs (Roder et al. 1998) and indicates the particular value of SSRs for comparative mapping and the identification of homoeologues in allopolyploid species. In addition to the homoeologous loci detected, a number of white clover SSRs which mapped to only one homoeologue produced other amplicons in the same parent that did not segregate and which may be monomorphic homoeoloci. This agrees with observations using RFLP probes in allotetraploid cotton mapping populations, where numerous dominant/recessive marker phenotypes corresponding with an intense signal at monomorphic loci suggests the presence of monomorphic homoeoloci in the other subgenome (Reinisch et al. 1994).

Using 365 SSRs, we were able to develop a medium-density genetic map (mean 2.3 cM per locus) of the allotetraploid white clover genome. Exclusive reliance on SSR data, and primarily EST-SSRs, is perceived to cause genome coverage to be biased relative to anonymous marker technologies such as genomic RFLP or random amplified polymorphic DNA (RAPD) (Li et al. 2002; Morgante et al. 2002). Our dataset exhibited variable distribution of SSR loci within the eight homoeologous pairs, between the homoeologues themselves and within each homologue. Despite a marker bias toward homoeologous pairs A, B, and D, we were able to identify markers on all 16 linkage groups. The density ranged from 1.6 cM per locus on groups B1 and D1 to 4.2 cM per locus on group F2. This inter-group bias is similar but less extreme than the bias observed in the low-density map of the white clover genome developed using AFLP and SSR marker systems, where 45% of the loci detected were on just four of the 18 linkage groups (Jones et al. 2003). We did not observe biased genetic distribution of genomic SSRs relative to EST-SSRs, which concurs with the evidence for interspersion of marker loci developed from these two sources in a previously reported comprehensive SSR map of the maize genome (Sharopova et al. 2002).

Homoeologues within pairs exhibited similar marker densities, with pair C providing a notable exception. Although the progenitor species and subgenome identities have not been conclusively resolved for white clover, a maximum subgenome locus difference may be estimated by assuming that the group within each homoeologous pair with the least loci belongs to the same subgenome.

This calculation indicates that for this dataset the maximum difference between the two subgenomes in number of loci detected is 20% (273 vs. 224). This narrow maximum difference suggests that there has been no general relaxation of selection pressure in one subgenome due to wholesale buffering by the other subgenome. However, based on this limited dataset, and without an understanding of ancestry and the evolutionary history of white clover, we can only speculate and cannot draw any conclusion regarding the selection pressures or the fate of particular homologues, chromosome regions or specific loci within genomes. Recent work in allotetraploid cotton and the two diploid progenitor genomes has documented the continued functionality and independent evolution of both subgenomes at the tetraploid level, with organ-specific specialisation of expression for some loci (Adams et al. 2003; Cronn et al. 1999). Furthermore, these authors determined that for the genes sampled, pseudogenisation was the exception rather than the rule. Assuming that pseudogenes would mutate at a faster rate than active open reading frames, our data demonstrate a general homogeneity of EST-SSR polymorphism levels within homoeologue pairs and suggests that both subgenomes of white clover are still functional and have evolved at relatively similar rates.

In accordance with the observations made regarding loci density, we did not observe substantial differences in map length within or between homoeologous pairs, with most of the differences in length being either marginal or easily attributed to variable marker density. This is in broad agreement with the homogeneous karyotype of white clover. Within homoeologous pairs, if we assume that each longer linkage group of the pair belongs to the same subgenome and that G2 is average, we can estimate a maximum genetic length difference of 10% (624 cM vs. 563 cM) between the two subgenomes of allotetraploid white clover. When considered in tandem with the same density distribution, we again see evidence for two relatively similar subgenomes in the white clover nucleus.

Marker density within a linkage group was often biased toward the distal ends of linkage groups (most evident in B2). This bias has been observed in other eukaryotes; for example, rice EST density has been shown to vary across genetic and physical distance, with distal maxima on chromosomes (Wu et al. 2002). Despite the density trends often observed within and among linkage groups, EST maps have been developed which give a high-density comprehensive coverage of well-characterised genomes including *Arabidopsis* (Lan et al. 2000), rice (Wu et al. 2002) and maize (Davis et al. 1999). Taking these reports into account, and given the genetic distribution of white clover EST-SSRs observed in the present work, we suggest that EST-SSRs will provide comprehensive coverage throughout both subgenomes of white clover and that with the implementation of predictive bioinformatic tools we can rapidly fill gaps and increase marker density in targeted regions of the white clover genome. We cannot, however, assume that all physical regions of the white clover genome will be tractable for EST-based marker systems.

With the exception of pair G, comparisons of locus order within homoeologous pairs suggest general colinearity of the two subgenomes of white clover, which may indicate a high level of propinquity between the progenitor genomes and either a relatively recent polyploidisation event or systematic preservation of colinearity, as we did not detect the chromosomal aberrations often detected in other allopolyploid species (Reinisch et al. 1994). Although deviations from homosequential ordering were observed in many instances, these often occurred in regions of high marker density bordering on regions of low marker density in one or both homoeologues (e.g. *prs431x* on B1 and *prs431a* on B2), suggesting that filling the remaining genetic space and/or observing additional meioses may resolve many of these minor discrepancies. The proportion (7%) of SSRs which detected duplicate loci either within homologues or among homoeologous pairs is higher than that observed in other white clover mapping studies (Jones et al. 2003). These non-syntenic loci may be attributed to either non-specific priming, paleohomoeology within or between subgenomes or the association of SSRs with mobile elements such as has been observed in the barley genome (Ramsey et al. 1999).

This map of the white clover genome used a double-pseudo testcross structure, a design that has been reported to minimise the impact of genetic load and self-infertility on segregation distortion and make the best use of highly heterozygous genotypes (Grattapaglia and Sederoff 1994). Of the loci observed in the 364/7×6525/5 population, 6% exhibited segregation distortion in the progeny sample, which compares favourably to the 11% detected in a white clover F_2 population based on a pair-cross between two partial inbreds carrying the S^f (self-fertile) allele (Jones et al. 2003) and the 37% detected in a *T. pratense* backcross population (Isobe et al. 2003), and is similar to that reported in backcross populations of the autotetraploid *M. sativa* (Brouwer and Osborn 1999). Substantial segregation distortion has also been detected in the self-fertile species *M. truncatula*, with areas on groups 1, 2, and 3 exhibiting high levels of distortion (Thoquet et al. 2002). A clear positional trend for segregation distortion was also observed in this white clover dataset, with three homoeologous groups (A, D, H) exhibiting clusters of distorted markers. As observed in *M. truncatula*, both white clover genotypes produced similar numbers of distorted loci, although the significant area of distortion on D2, which prevented the ordering of 12 loci on that group, originated only from parent 6525/5. These patterns of segregation distortion observed in *M. truncatula* and white clover contrast with the random distribution and parental bias of marker distortion observed in the red clover genetic map (Isobe et al. 2003). Collectively, these data suggest that there is a substantial genetic load in some species of Trifolieae and that this may lead to segregation distortion, which may in turn be removed by the selection of appropriate parental material and population architecture.

The total map length of 1,144 cM is a preliminary estimate due to G2 not coalescing in the bi-parental consensus analysis; however, it still allows for comparison with other Trifolieae maps. A white clover map calculated in MAPMAKER 3.0 from an F_2 dataset was 825 cM (Jones et al. 2003), just over the theoretical minimum of one chiasma per bivalent. However, that report was almost certainly a substantial underestimate of the genetic size of the white clover genome, given that the map did not coalesce to 16 groups, only four of the 18 observed linkage groups had greater than ten loci and no homoeologous pairs were identified. In contrast, the map we report is of a higher density (2.3 cM per locus) and, with the exception of G2, is likely to be representative of the white clover genome (Fig. 4). With an estimated length of 1,144 cM and assuming that G2 is genetically average (given the 16 homologues are morphologically similar), a total length of approximately 1,200 cM is probable using JOINMAP in an F_1 population. JOINMAP has been documented to produce shorter maps than MAPMAKER when interference is present (Qi et al. 1996), and re-analysis of single-parent linkage groups in our dataset using MAPMAKER 3.0 has resulted in approximately a 30% increase in map length (data not shown). Notably, the estimate of 1,200-cM total map length in tetraploid clover is, on a per ploidy basis, in close agreement with the estimate of the genetic length of the sequence-based map of the diploid *M. truncatula* genome at approximately 600 cM (Choi et al. 2004).

Division of the estimated physical size of the white clover genome (956 Mbp) by the estimated map length of 1,144 cM results in a ratio of 0.84 Mbp/cM, a figure that closely approximates the ratio reported in *M. truncatula* at the *sunn* locus (Schnabel et al. 2003) but is higher than that around the *DMI* loci (Ané et al. 2002). This demonstrates that white clover has a very compact allopolyploid genome that is well suited for genetic analysis. Localisation of the *R* locus, which confers a spectrum of anthocyanin accumulator phenotypes in white clover leaves, onto group B1 in close genetic proximity (1.1 cM) to locus *prs318* is an illustration of the usefulness of this population for the genetic mapping of key traits in white clover. The ratio of 0.84 Mbp/cM suggests the loci may be separated by less than 1 Mbp, however, implementation of fine mapping strategies will be necessary to refine this estimate.

Given the compact allotetraploid genome (twice the size of *M. truncatula*), the similarity in genetic and physical size of the two subgenomes and the capability to develop synthetic allotetraploids (W. Williams, personal communication), white clover is well suited for the investigation of gene evolution and expression in allotetraploids (Adams et al. 2003; Jiang et al. 1998). However, it will be necessary to determine the evolutionary origin(s) of white clover and define an ancestral genome nomenclature within *Trifolium* prior to pursuing this in any detail. Given the genome specificity of many of the marker loci mapped this study, it may be possible to use genetic markers alongside DNA sequence-based molecular phylogeny to establish a robust ancestry of allotetraploid white clover and to identify its geographic centre(s) of origin.

This paper describes the first large-scale implementation of EST-SSRs for genetic linkage analysis in an under-characterised allopolyploid genome. Given the high frequency of homoeologous loci detection, we can conclude that SSRs sourced from transcribed regions of the genome are a particularly useful resource of PCR-based molecular markers. This medium-density map of the clover genome will provide a springboard for the dissection of quantitative traits into Mendelian characters, marker-assisted breeding, comparative genome analysis, investigation of white clover genome putative ancestry and the positional cloning of key white clover genes involved in sustainable forage production.

Acknowledgements The authors recognise the technical contributions of Greig Cousins, Kenyon Moore, Benjamin Franzmayr and Deborah Knox. Funds were provided by the New Zealand Foundation for Research, Science, and Technology under contract no. C10X0203. Mapped and novel EST-SSRs are available via Material Transfer Agreement for research purposes or via license for commercial purposes.

References

Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci USA 100:4649–4654

Ane J, Levy J, Thoquet P, Klikova O, de Billy F, Penmetsa V, Kim D, Debelle F, Rosenburg C, Cook D, Bisseling T, Huguet T, Dearie J (2002) Genetic and cytogenetic mapping of *DMI1*, *DMI2* and *DMI3* genes of *Medicago truncatula* involved in Nod factor transduction, nodulation, and mycorrhization. Mol Plant Microbe Interact 15:1108–1118

Areshchenkova T, Ganal M (2002) Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. Theor Appl Genet 104:229–235

Atwood SS (1940) Genetics of cross-incompatibility among self-incompatible plants of *Trifolium repens*. J Am Soc Agron 32:955–968

Atwood SS, Hill HD (1940) The regularity of meiosis in microsporocytes of *Trifolium repens*. Am J Bot 27:730–735

Atwood SS, Sullivan JT (1943) Inheritance of a cyanogenic glucoside and its hydrolysing enzyme in *Trifolium repens*. J Hered 34:311–320

Awadalla P, Ritland K (1997) Microsatellite variation and evolution in the *Mimulus guttatus* species complex with a contrasting mating system. Mol Biol Evol 14:1023–1034

Badr A, Sayed-Ahmed H, El-Shanshouri A, Watson LE (2002) Ancestors of white clover (*Trifolium repens* L.) as revealed by isozyme polymorphisms. Theor Appl Genet 106:143–148

Band MR, Larson JH, Rebeiz M, Green CA, Heyen DW, Donovan J, Windish R, Steining C, Mahyuddin P, Womack JE, Lewin HA (2000) An ordered comparative map of the cattle and human genomes. Genome Res 10:1359–68

Bennett MD, Leitch IJ (2003) Angiosperm DNA C-values database (release 4.0, January 2003) http://www.rbgkew.org.uk/cval/homepage.html

Brouwer DJ, Osborn TC (1999) A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). Theor Appl Genet 99:1194–1200

Brownstein M, Carpten J, Smith J (1996) Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. BioTechniques 20:1004–1010

Caradus JR, Clifford PTP, Chapman DF, Cousins GR, Williams WM, Miller JE (1997) Breeding and description of grasslands sustain. N Z J Agric Res 40:1–7

Carnahan HL, Hill HD, Hanson AA, Brown KG (1955) Inheritance and frequencies of leaf markings in white clover. J Hered 46:109–114

Chen CC, Gibson PB (1970) Meiosis in two species of *Trifolium* and their hybrids. Crop Sci 10:188–189

Cho Y, Ishii T, Temnykg S, Chen X, Lipovich L, McCouch S, Park W, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa*). Theor Appl Genet 100:713–722

Choi HK, Kim D, Uhm T, Limpens E, Lim H, Kalo P, Penmetsa R, Seres A, Kulikova O, Bisseling T, Kiss GB, Cook D (2004) A sequence-based map of *Medicago truncatula* and comparison of marker co-linearity with *Medicago sativa*. Genetics 166 (in press)

Corkill L (1971) Leaf markings in white clover. J Hered 62:307–310

Cronn RC, Small RL, Wendel JF (1999) Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci USA 96:14406–14411

Davis GL, McMullen MD, Baysdorfer C, Musket T, Grant D, Staebell M, Xu G, Polacco M, Koster L, Melia-Hancock S, Houchins K, Chao S, Coe EH Jr (1999) A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged sites (ESTs) in a 1736-locus map. Genetics 152:1137–1172

Decroocq V, Favé M, Hagen L, Bordenave L, Decroocq S (2002) Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor Appl Genet 106:912–922

Dvorak J, Luo M-C, Yang Z-L (1997) Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics 148:423–434

Eujayl I, Sorrells M, Baum M, Wolters P (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Eujayl I, Sledge M, Wang L, May G, Chekhovskiy K, Zwonitzer JC, Mian M (2004) *Medicago truncatula* EST-SSRs reveal cross species genetic markers for *Medicago* spp. Theor Appl Genet 108:414–422

Fulton T, Van der Hoeven R, Eannetta N, Tanksley S (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Gillett JM (1985) Taxonomy and morphology. In: Taylor NL (ed) Clover science and technology. American Society of Agronomy, Madison, Wis., pp 7–69

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Griffiths AG, Barrett BA, Ellison N, Sawbridge T, Spangenburg G, Bryan GJ, Schreiber M (2002) Getting out there: transportability of white clover EST and genomic DNA-derived microsatellites. In: McComb J (ed) Proc 12th Australasian Plant Breed Conf. Perth, Australia, pp 69–72

Gualtieri G, Kulikova O, Limpens E, Kim DJ, Cook DR, Bisselin T, Geurts R (2002) Microsynteny between pea and *Medicago truncatula* in the *SYM2* region. Plant Mol Biol 50:225–235

He C, Poysa V, Yu K (2003) Development and characterization of simple sequence repeat (SSR) markers and their use in determining relationships among *Lycopersicon esculentum* cultivars. Theor Appl Genet 106:363–373

Holton T, Christopher J, McClure L, Harker N, Henry R (2002) Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat. Mol Breed 9:63–71

Isobe S, Klimenko I, Ivashuta S, Gau M, Kozlov N (2003) First RFLP linkage map of red clover (*Trifolium pratense* L.) based on cDNA probes and its transferability to other red clover germplasm. Theor Appl Genet 108:105–112

Jiang C, Wright R, El-Zik K, Paterson AH (1998) Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proc Natl Acad Sci USA 95:4419–4424

Jones E, Hughes L, Drayton M, Abberton M, Michaelson-Yeats T, Bowen C, Forster J (2003) An SSR and AFLP molecular marker-based genetic map of white clover (*Trifolium repens* L.). Plant Sci 165:531–539

Julier B, Flajoulot S, Barre P, Cardinet G, Santoni S, Huguet T, Huyghe C (2003) Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. BMC Plant Biol 3:9

Kalo P, Endre G, Zimanyi L, Csanadi G, Kiss GB (2000) Construction of an improved linkage map of diploid alfalfa (*Medicago sativa*). Theor Appl Genet 100:641–657

Kantety RV, La Rota M, Matthews DE, Sorrells M (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510

Kolliker R, Jones ES, Drayton MC, Dupal MP, Forster JW (2001) Development and characterisation of simple sequence repeat (SSR) markers for white clover (*Trifolium repens* L.). Theor Appl Genet 102:416–424

Lan TH, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, Kresovich S, Paterson AH (2000) An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res 10:776–788

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11:2453–2465

Matthews BF, Devine TE, Weisemann JM, Beard HS, Lewers KS, MacDonald MH, Park Y-B, Maiti R, Lin J-J, Kuo J, Pedroni MJ, Cregan PB, Saunders JA (2001) Incorporation of sequenced cDNA and genomic markers into the soybean genetic map. Crop Sci 41:516–521

Mercer CF, van den Bosch J, Miller K (1999) The effectiveness of recurrent selection of white clover (*Trifolium repens*) for resistance to New Zealand populations of clover cyst nematode (*Heterodera trifolii*). Nematology 1:449–455

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genet Res 10:72–80

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Qi X, Stam P, Lindhout P (1996) Comparison and integration of four barley genetic maps. Genome 39:379–394

Ramsey L, Macaulay M, Cardle L, Morgante M, Degli Ianissevich S, Maestri E, Powell W, Waugh R (1999) Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J 17:415–425

Reinisch A, Dong J, Brubaker CL, Stelly D, Wendel JF, Paterson AH (1994) A detailed RFLP map of Cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics 138:829–847

Roder M, Korzun V, Wendchake K, Plaschke J, Tixier M, Leroy P, Ganal M (1998) A microsatellite map of wheat. Genetics 149:2007–2023

Sanwen H, Baoxi Z, Milbourne D, Cardle L, Guimei Y, Jiazhen G (2000) Development of pepper SSR markers from sequence databases. Euphytica 117:163–167

Sawbridge T, Ong E, Binnion C, Emmerling M, Meath K, Nunan K, O'Neill M, O'Toole F, Simmounds J, Wearne K, Winkworth A, Spangenburg G (2003) Generation and analysis of expressed sequence tags in white clover (*Trifolium repens* L.). Plant Sci 165:1077–1087

Schnabel E, Kulikova O, Penmetsa R, Bisseling T, Cook D, Frugoli J (2003) An integrated physical, genetic and cytogenetic map around the *sunn* locus of *Medicago truncatula*. Genome 46:665–672

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100723–726

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, Liscum E, Cone K, Davis G, Coe EH Jr (2002) Development and mapping of SSR markers for maize. Plant Mol Biol 48:463–81

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M, Powell W (2003) How much effort is required to isolate nuclear microsatellites from plants? Mol Ecol 12:1339–1348

Tasma IM, Shoemaker RC (2003) Mapping flowering time gene homologs in soybean and their association with maturity (E) loci. Crop Sci 43:319–328

Temnykh S, DeClerck G, Lukashova A, Leonard L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations and genetic marker potential. Genet Res 11:1441–1452

Thoquet P, Gherardi M, Journet EP, Kereszt A, Ane JM, Prosperi JM, Huguet T (2002) The molecular genetic linkage map of the model legume *Medicago truncatula*: an essential tool for comparative legume genomics and the isolation of agronomically important genes. BMC Plant Biol 2:1

Thornsberry J, Goodman M, Doebley J, Kresovich S, Nielsen D, Buckler E (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J (2002) Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol 19:1251–1260

Williams WM (1987) White clover taxonomy and biosystematics. In: Williams WM (ed) White clover. CAB Int, Oxon, pp 323–342

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14:525–535

Zhu H, Cannon S, Young N, Cook D (2002) Phylogeny and genomic organization of the TIR and non-TIR NBS-LRR Resistance Gene Family in *Medicago truncatula*. Mol Plant Microbe Interact 15:529–539

Zhu H, Kim DJ, Baek JM, Choi HK, Ellis LC, Kuester H, McCombie WR, Peng HM, Cook DR (2003) Syntenic relationships between *Medicago truncatula* and *Arabadopsis* reveal extensive divergence of genome organization. Plant Physiol 131:1018–1026