Experimental estimation of mutation rates in a wheat population with gene genealogy approach

Anne-Laure RAQUIN(*) , Frantz DEPAULIS(†), Amaury LAMBERT(†), Nathalie GALIC(*), Philippe BRABANT(*), Isabelle GOLDRINGER(*).

(*) UMR de Génétique Végétale, INRA, CNRS, Univ Paris Sud, AgroParisTech, Ferme du Moulon, F91190 GIF-SUR-YVETTE, FRANCE.

(†) Laboratoire d’Ecologie, CNRS, UMR 7625, Ecole Normale Supérieure, 75230 PARIS Cedex 05, FRANCE.
Running head: Mutation rate estimation

Key words: microsatellites, genetic diversity, mutation rate, gene genealogy, Triticum aestivum L.

Corresponding Author: Isabelle Goldringer
UMR de Génétique Végétale, INRA/CNRS/Univ Paris Sud/AgroParisTech
Ferme du Moulon
91190 GIF-SUR-YVETTE
Tel: 33 1 69 33 23 70
Fax: 33 1 69 33 23 40
isa@moulon.inra.fr
ABSTRACT

Microsatellite markers are extensively used to evaluate genetic diversity in natural or experimental evolving populations. Their high degree of polymorphism reflects their high mutation rates. Estimates of the mutation rates are therefore necessary when characterizing diversity in populations. As a complement to the classical experimental designs, we propose to use experimental populations, where the initial state is entirely known and some thoroughly surveyed intermediate states, thus providing short time scale estimation together with a large number of cumulated meioses. In this article, we derived four original gene genealogy based methods to assess mutation rates with limited bias due to relevant model assumptions incorporating the initial state, the number of new alleles and the genetic effective population size. We studied the evolution of genetic diversity at 21 microsatellite markers, after 15 generations in an experimental wheat population. Compared to the parents, 23 new alleles were found in generation 15 at nine out of the 21 loci studied. We provide evidence that they arose by mutation. Corresponding estimates of the mutation rates ranged from 0 to 4.97 x 10^{-3} per generation (i.e. year). Sequences of several alleles revealed that length polymorphism was only due to variation in the core of the microsatellite. Among different microsatellite characteristics, both the motif repeat number and an independant estimation of the Nei diversity were correlated with the novel diversity. Despite a reduced genetic effective size, global diversity at microsatellite markers increased in this population, suggesting that microsatellite diversity should be used with caution as an indicator in biodiversity conservation issues.
INTRODUCTION

Because microsatellite markers (tandemly repeated DNA motifs of 1-6 bp in length) are highly polymorphic and since they are distributed across the whole genome (Pejic et al. 1998; Plaschke et al. 1995; Wu and Tanksley 1993), they constitute a powerful tool to assess the level of genetic differentiation within and among experimental or natural populations at different generations. The high degree of polymorphism at microsatellite markers is directly related to their underlying mutation rates, which can be explained by two mutational mechanisms: polymerase slippage during DNA replication and unequal crossing-over during recombination—but not excluding SNP mutations at a lower rate. These two mechanisms involve changes in the number of motif repeats. Understanding the evolutionary properties of microsatellites is hence necessary for correctly interpreting diversity data when studying populations across generations and/or populations which have spatially diverged (Ellegren, 2004).

An increasing number of studies have been devoted to the estimation of mutation rates at microsatellite loci (e.g. Schug et al. 1998; Denver et al. 2004; Symonds and Lloyd 2003; Thuillet et al. 2004), which reveal a far more complex scheme for microsatellite evolution than previously stated (Schlötterer 2000; Ellegren 2004). Parallel to the empirical studies, more refined and diverse mutation models of microsatellite evolution have been developed such as the proportional slippage (PS) model of Kruglyak et al. 1998 (where length mutations tend to increase with increasing repeat numbers, balanced by point mutations breaking the longer sequences into smaller units) and its derivatives (e.g. Xu et al. 2000). To provide valuable data for testing the different models in a large number of species, data acquisition designs should allow for unbiased and accurate estimations of mutation rates. As far as we know, four kinds of approaches have been used to estimate mutation rates in a wide range of species. They can be described according to the time scale they consider: a short time scale
when using (i) pedigree based (or parent-offspring) estimation (in humans: Weber & Wong 1993; Heyer et al. 1997; Whittaker et al. 2003, birds: Brohede et al. 2002; Beck et al. 2003, Brohede et al. 2004 and in Gastropoda: Gow et al. 2005) or (ii) mutation accumulation lines (Vigouroux et al. 2002 in maize; Thuillet et al. 2002 in wheat; Schug et al. 1998 and Vazquez et al. 2000 in drosophila; Denver et al. 2004 and Seyfert et al. 2008 in Caenorhabditis elegans; Mc Connell et al. 2007 in Dictyostelium discoideum), and a long time scale with (iii) inter-species comparison scaling divergence time with fossil data (human/chimp in Webster et al. 2002 and Sainudiin et al. 2004; Kayser et al. 2006) or (iv) within-species allele frequency distribution (Thuillet et al. 2004 in wheat; Chakraborty et al. 1997 and Xu et al. 2005 in humans; Symonds et al. 2003 in Arabidopsis thaliana). Short time scale approaches (pedigree or accumulation lines) are expected to be less biased since they rely on the direct observation of neo-mutations but in order to reach a sufficient accuracy they need larger sample sizes to allow for the detection of rather rare events. Long time scale approaches take advantage of the numerous meiôses accumulated since the time when populations or species under comparison diverged, but they suffer from the potential confounding effects of selection, demography and risks of saturation, which might lead to biased estimates (Zhivotovsky et al. 2006).

While these studies are at least in agreement as to the finding of high heterogeneity in mutation rates with respect to loci and on the positive correlation of mutation rate or diversity with microsatellite size or repeat number, they led to contrasting results regarding the number of repeats of a mutation (stepwise mutation model -SMM- vs two-phase model -TPM) and the occurrence of a bias in the direction of the mutation (increasing or decreasing allele size).

Moreover, the estimated mutation rates seemed to vary depending on the experimental design used: the estimated mutation rates were lower in long-scale designs (human-chimp comparison, Webster et al. 2002; Sainudiin et al. 2004) than in pedigree / parent-offspring
designs (Weber & Wong 1993; Heyer et al. 1997; Whittaker et al. 2003) in agreement with the saturation effect.

In this article, we develop a complementary approach to the above cited designs which involves an experimental population where the initial state is entirely known and some thoroughly surveyed intermediate states thus providing short time scale estimation together with a reasonable number of cumulated meïoses. In addition to limiting bias, this approach allows for estimating the genetic effective population size from temporal variation in allele frequencies (Waples 1989) at each locus providing indications of genetic drift and selection effect together with mutation rate at the individual locus level. Yet, assessing mutation rates from genetic samples taken over time within such evolving populations is not straightforward since the genealogy between the different samples is usually not known. We therefore derived four original gene genealogy based methods to assess mutation rates with limited bias due to relevant model assumptions incorporating the initial state, the number of new alleles and the genetic effective population size. Genetic data at microsatellite markers were collected from an experimental wheat population involved in 15 generations of evolution under semi natural/controled conditions. We identified the new diversity, studied its dynamics over generations and assessed mutation rates using the newly developed methods. These results provide new insights into the dynamics, maintenance, and renewal of microsatellite variability in an experimental evolving population. We also address the issue of using microsatellite markers with high mutation rates to make inferences on the evolution and conservation of genetic diversity within populations, at a time scale relevant for the conservation of genetic resources.
MATERIALS AND METHODS

Population studied: The wheat experimental population was part of a dynamic management program aimed at investigating an evolutionary conservation method for the genetic resources of cultivated species (Goldringer et al. 2006). Sixteen homozygous genotypes were used as the parents of the population. These 16 genotypes were manually crossed by pairs; the eight resulting hybrids were then crossed again for three generations until obtaining a single segregating population where each individual was derived from all 16 parents with expected equal contributions (Figure 1). After the four generations of crosses, three generations of bulk multiplication (under the naturally mainly selfing mating system of wheat) were performed to increase the number of seeds of this population. This initial experimental population corresponds to generation 0 (hereafter G0). In 1984, G0 seeds were sown at Le Moulon (48°4’N 2°1’E, near Paris) where the population has then been cultivated for 16 generations (one generation / year from mid-1984 to mid-2000). From 5,000 to 10,000 plants were cultivated each generation in a 100 m² plot isolated from other wheat fields to avoid cross-pollination, both by distance (>50m) and by the physical barrier of a different and taller species (rye,...) surrounding the plot. At each generation, all the seeds were bulk harvested and mixed, and a sample of 10,000 seeds was sown providing 5,000 to 10,000 plants at the next generation. Temporal evolution of microsatellite diversity was studied among the 16 parents and samples of individuals from different generations. In spring 2000, 287 individuals of generation 15 (G15: plants grown from seeds harvested after 15 generations) were randomly sampled within the population to be genotyped. Seeds harvested after 1 (G1), 5 (G5) and 10 (G10) generations were conserved in a cold room and a sample of them (159, 170 and 130 individuals respectively) was grown in the field for the study. The 16 parents were conserved in a cold room at the INRA laboratory Le Moulon, and were multiplied at maximum four times by self-pollination for maintaining seed germination ability.
**Molecular analysis:** Total DNA of the 16 parents and of the 287 G15 individuals was extracted from 200 mg of young leaves following a CTAB protocol adapted from Murigneux *et al.* (1993). Nineteen monolocus microsatellite markers developed by Röder *et al.* (1998) (*Xgwm* markers from IPK Gatersleben), and one microsatellite marker (*Cfd71*) developed at the INRA Clermont-Ferrand laboratory (Guyomarc'h *et al.* 2002a) amplifying two loci (*Cfd71_A* and *Cfd71_D*), were chosen to be polymorphic within the set of parents and to cover all 21 chromosomes of the wheat genome. The 21 microsatellites were amplified by PCR from 50 ng genomic DNA in 25 µl volumes using 1.5 mM MgCl₂, 200 µM dNTPs, 1 µM of each primer, and 1 U *Taq* polymerase. Cycling conditions were the same as defined by Röder *et al.* (1998), except for *Xgwm261*, where the annealing temperature was set to 60°C. PCR products were separated by electrophoresis in 6% sequencing polyacrylamide gels, visualized by silver staining according to Tixier *et al.* (1997) and sized using a 10bp DNA ladder. DNA of five parents polymorphic for the microsatellite loci were used as references in each run. PCR, electrophoresis and silver staining were realized at the INRA Clermont-Ferrand.

At each locus, the different alleles found among the 16 parents (parental alleles) were identified and characterized by their size (in bp). In G15, all alleles at each locus were compared to the parental alleles, and novel alleles (not present among the parents) were identified.

To analyze more precisely the emergence and evolution of these new alleles in the population, we genotyped individuals sampled at G1, G5 and G10 together with the 16 parents for six loci characterized by new alleles at G15. Total DNA of these individuals was extracted from lyophilized young leaves following a rapid procedure adapted from Dellaporta *et al.* (1983). Forward primers were modified with an M13 extension according to Boutin-Ganache *et al.* (2001). The microsatellites were amplified from 200 ng genomic DNA in 10 µl volumes.
using 1.5 or 3 mM MgCl2, 200 µM dNTPs, 25 nM fluorescent labeled M13 tail, 0.5 µM of each modified forward primer, 0.5µM of each reverse primer and 1 U Taq polymerase.

Cycling conditions were 5 min at 95°C, 30 cycles of 20 s at 95°C, 20 s at 50-60°C (depending on the primer, according to Röder et al., 1998 and Guyomarc’h et al., 2002a), 30 s at 72°C, and 3 min at 72°C. The PCR products were separated by electrophoresis in 6.5% sequencing polyacrylamide gels in a LiCor automated sequencer (LiCor Biosciences). The amplified fragments were analysed with the version 2.03 of the OneDscan software (Scanalytics). To compare alleles identified from both methods (electrophoresis with silver staining and electrophoresis on LiCor automated sequencer), all 16 parental alleles were resized with this second method.

**Cloning and sequencing:** To investigate the nature of length variation in microsatellite polymorphism and to determine whether the allelic diversity observed was due to changes in the number of repeats or in the flanking sequences, some new alleles together with the parental ones were sequenced. For one locus with few alleles, we cloned all alleles before sequencing. For the others, because of their large number of alleles, we analysed only a sample of alleles as described in the following:

a) we cloned one allele per locus before sequencing to establish a reference sequence,

b) we sequenced in both ways the other alleles directly from PCR products.

First, all alleles were amplified by PCR with unlabelled primers, as described in the above section, and the fragments were separated by electrophoresis on 6% denaturing polyacrylamide gels containing 7.5M urea, 6% acrylamide, and 1X TBE buffer. PCR products were visualized by silver staining according to Tixier et al. (1997) and sized with a 10pb DNA ladder. DNA bands with the appropriate size (corresponding to the most intense signal) were extracted and purified with the QiaexII kit (Qiagen). To increase the concentration of microsatellite fragments (loss of material linked to the steps of band purification) and hence
the probability of success for cloning and sequencing, we amplified DNA fragments by another PCR using the same conditions as the initial one. After this step, one allele per locus was cloned before sequencing (Method A), and the others were directly sequenced in both directions at the Montpellier INRA laboratory on a ABI377 semi-automated sequencer (Method B). In Method A, the reference allele was first ligated into pGEM-T vector (Promega) and then cloned into Escherichia coli DH5α chemically competent cells (Invitrogen). Because PCR generated stutter (small noise bands close to the allele one), we cloned the allele with its potential stutter. To assess the exact size of each insert and to select colonies carrying the right allele, white colonies were amplified by PCR using labelled M13 primers (Boutin-Ganache et al. 2001), and PCR products were separated on a LiCor automated sequencer with a parental control. The colonies carrying the target insert were sequenced at Genomexpress. The flanking sequences obtained with Method B) were compared with those obtained in the reference sequence (Method A). While the sequencing of the microsatellite alleles did not provide their exact size (number of repeats), it allowed us to determine the nature of the observed polymorphism (in the core of the microsatellite or in the neighbouring regions).

**Diversity analysis:** At each locus, gene diversity (Nei, 1973) was calculated in the parents and its unbiased estimate (Nei 1987) was used for G15 to account for sampling variation on the allele frequency estimations.

Genetic effective population size ($N_{eg}$) between parents and G15 was estimated based on temporal changes in microsatellite allele frequencies according to Waples (1989):

$$
\hat{N}_{eg} = \frac{t}{2\left(\hat{F}_c - \frac{1}{S_i}\right)}
$$

(1)

where $\hat{F}_c$ is the multilocus estimate of the standardized variance of allelic frequencies (Nei and Tajima 1981), $t$ is the number of generations between the two populations (there were 22
generations between the parents and G15) and $S_t$ the sample size at the final generation. Sampling error was ignored in the 16 parents because they allowed the calculation of the exact initial allele frequencies. A confidence interval (CI) for $\hat{N}_{eg}$ was derived based on the Chi-square approximation for $n\hat{F}_c E(\hat{F}_c) = \text{number of alleles–number of loci}$ (Waples 1989).

Genetic effective population size estimated from marker data was compared to the demographic effective size $N_{ed}$ estimated as $N/(1+ F_{is})$ (Caballero 1994) where $N$ is an estimator of the census population size (an underestimate of 5,000 individuals based on the minimum number of plants grown at each generation was considered here), and $F_{is}$ was the average inbreeding coefficient calculated on the 21 loci in a population at inbreeding equilibrium. If a population is only faced with genetic drift, the demographic and the genetic effective population sizes should be the same, i.e. $N_{eg} \approx N_{ed}$. To identify markers exhibiting extreme $F_c$-values compared to the rest of the genome, we tested if the temporal allelic variation observed at each locus was significantly higher than expected under genetic drift alone. To test the null hypothesis “temporal allelic variation is homogeneous throughout the genome”, each $F_c$-value was compared to the distribution obtained from 3,000 random independent simulation runs (Goldringer and Bataillon 2004) based on the mean genetic effective population size, $N_{eg}$, estimated between the parents and G15. Simulations were carried out using the Mathematica software (Wolfram 1996).

**Testing for migration:** To determine the origin of the novel diversity observed in the population, we identified sources of potential migration by listing all the field wheat varieties cultivated at Le Moulon since 1984 (supplemental Table 1) and all these varieties were genotyped at the Clermont-Ferrand INRA laboratory with the same 21 markers used for the parents and G15. Assuming that mutations occur independently at the different loci, we also tested for random association between new alleles at the 21 loci with a Chi-square test,
considering that at each locus, there were two classes of alleles: parental or new, and calculating the expected distribution of individuals with 0, 1, 2, ... 21 new alleles at the 21 loci under random association of new alleles. Whereas departure from random association would provide evidence for a migration origin of most new alleles, a non significant test could both indicate independent appearance by mutation and migration with migrants genetically related to the population.

**Mutation rates estimation:** The gene genealogy mutation rate estimators are described and discussed in the appendix. These are moment (mean) estimators relying on the computation of the size of the partial tree along the experiment (proportional- with a factor $1/\mu$- to the observed number of mutations). In short, the first two estimators are obtained under Kingman’s coalescent model assuming continuous time and a binary tree, neglecting multiple common ancestry events occurring at the same generation. The first estimator $\mu_1$ is computed with an approximation replacing coalescence times by their expectations and neglecting the dependence between them induced by the conditioning on an absolute duration of the experiment. The second one, $\mu_2$ is an exact analytical formula giving the size of the tree taking this dependence into account. The estimators $\mu_3$ and $\mu_4$ were derived using the Wright-Fisher model (WF, discrete generations, allowing for multiple simultaneous common ancestry events). An approximation formula which computed time recursively and replaced the number of ancestors by its expected value at each generation was used for $\mu_3$, and a Monte-Carlo backward simulation algorithm of gene genealogy associated to the WF model was used for $\mu_4$. The results related to the difference between the estimators are described and discussed in the appendix. In short, the two estimators provided fairly consistent estimates.

**Factors influencing locus diversity and mutation:** To obtain more insight into factors that influence the origin and renewal of genetic diversity, we searched for associations between characteristics of each of the 21 microsatellite loci and the number of parental and novel
alleles the analysis revealed. The characteristics we investigated were: the motif (type, perfection and number of repeats); the position on the chromosome with regard to the centromere; the genome (A, B or D) the locus belongs to; and the genetic diversity measured on a large collection of wheat genetic resources (details are given in supplemental Table 2). Note that given the number of loci studied (21 among which nine with new alleles), the statistical power was low, suggesting that significant tests should reveal a great magnitude of the effects involved. The effects of motif type, motif perfection, and genome location were investigated factor by factor using an analysis of variance (procedure GLM, SAS, 1999). Effects of the number of repeats, the locus position and the genetic diversity were tested separately by regression (procedure GLM, SAS, 1999).
RESULTS

Evolution of microsatellite diversity within the wheat population: Among the parents, the number of alleles per locus ranged from two to eight, with an average of 4.1 alleles per locus. The number of alleles per locus ranged from two to ten in the G15 generation, with an average of 5.0 alleles per locus. Three out of 86 parental alleles were no longer found in G15, whereas 23 novel alleles were detected. Nei’s diversity index for each locus ranged from 0.117 to 0.805 in the parents and from 0.029 to 0.760 in the G15 generation. They thus appeared to be quite similar between the parents and G15 (Figure 2), except for four loci (Xgwm135, Xgwm149, Xgwm312 and Xgwm626) for which $H$ was smaller in G15.

Nine among the 21 loci showed novel alleles, and a total of 23 new alleles were observed in G15. Most of the new alleles were detected at low frequency (<0.05), except for three alleles at three different loci (Xgwm260, Xgwm642, Cfd71_D) with frequencies ranging from 0.069 to 0.185 in G15 (Table 1). Among the nine loci with new alleles in G15 (Table 1), six (Xgwm181, Xgwm260, Xgwm312, Xgwm437, Xgwm642 and Cfd71_D) were chosen to study emergence of novel diversity. At these six loci, a total of 20 new alleles were found at the different generations studied. Of these, 10 new alleles were found in G1 and were maintained in the following generations, two were observed for the first time in G5 and were maintained at low frequency in the population, one was found in G10 and G15, and four were found only in G15. The three other new alleles were observed only once in G1, G5 or G10. In G10, five alleles observed in the previous generations, and found again in G15, were not detected (Table 1). This lower allelic richness at this generation (G10) could be explained by the smaller sample size. Among the 20 new alleles observed at these six loci in the four generations, 12 were already present in G1, compared to eight others detected between G1 and G15 (Table 1). Among the 20 new alleles detected, 14 had a size very close to the parental allele size, differing by only (+/-) 1 microsatellite repeat (Table 1). Three new alleles
differed from parental alleles by (+/-) two repeats and the three other new alleles differed from parental alleles by three repeats or more.

The genetic effective population size $N_{eg}$ based on averaged $F_{c,l}$ over the 21 loci, estimated between the parents and the G15 generation ($N_{eg}=183$) was similar to the genetic effective population size estimated previously at different generations with RFLP markers (e.g. $N_{eg}=123$ between G0 and G10 and $N_{eg}=144$ between G1 and G10; Goldringer et al. 2001). Yet, it was very low compared to the demographic effective size ($N_{ef}=2,625$, with an estimated inbreeding coefficient of $F_u=0.893$ and the population census size taken as $N=5,000$ plants). Studying the individual $F_{c,l}$ values did not allow us to identify loci that would have locally experienced a reduced effective size compared to the rest of the genome ($N_{eg}=183$). When accounting for the multiple testing (21 loci) according to Storey and Tibshirani (2003), none of the loci was significant, even those exhibiting the highest $F_{c,l}$ values.

**Cloning and sequencing of alleles:** Seventeen alleles at five loci presenting a novel diversity were sequenced: one at locus Xgwm181, three at locus Xgwm312, two at locus Xgwm437, three at locus Xgwm642 and eight at locus Cfd71_D. Flanking sequences obtained directly from PCR products were compared to flanking sequences obtained from cloned alleles. The sequence at locus Xgwm181 was obtained from cloning and was compared to the sequence of *Triticum turgidum* subsp. *durum* available on the NCBI web site (Thuillet et al., 2005, accession number AY579595). The microsatellite flanking sequences were strictly conserved in both species, the only differences reflecting the number of unit repeats. At locus Xgwm312, sequences of the three alleles (one cloned and two from PCR products) were compared. The absence of polymorphism in flanking sequences led us to conclude that the observed polymorphism was in the core of the microsatellite. Sequences were then compared to a sequence of *Triticum turgidum* subsp. *durum* available on the NCBI web site (Thuillet et al.
The sequences of the two species only differed by a single base deletion in the microsatellite flanking sequence. At locus Xgwm437, the two sequences obtained (corresponding to one cloned allele and one allele sequenced from PCR product) were also compatible with a polymorphism in the microsatellite number of repeats. At locus Xgwm642, the three cloned sequences were compared to the sequence of *Aegilops tauschii* available on the NCBI web site (see supplemental Figure 1). The observed polymorphism among all alleles was systematically due to the number of microsatellite repeats. The flanking sequences were strictly conserved between the two species. At locus Cfd71_D, eight sequences (four parental and four new alleles) were obtained. Although no reference sequence was available for this locus, the polymorphism observed seemed to be also due to the number of repeats. All the new alleles sequenced in this study differed from parental alleles in the core of the microsatellite and polymorphism observed in cloned sequences was due to changes in the number of microsatellite motifs regardless of the species.

**Origin of new alleles:** Among the four bread wheat varieties (namely *Recital*, *Festival*, *Soissons* and *Theseep*), cultivated at Le Moulon since 1984 and identified as potential sources of migrants for the population, none of them contained any of the new alleles observed in G15 (see supplemental table 1 for more details).

Among the 287 individuals genotyped in G15, 176 had at least one new allele, but no individual had more than four new alleles simultaneously. The Chi-square test for independent associations of new alleles at different loci was not significant ($P = 0.43$) indicating that independence could not be rejected.

**Mutation rate estimation:** Mutation rates were estimated for each of the six loci studied at intermediate generations between parents and G1, and between G1 and G15, and were estimated for each of the nine loci with new alleles between parents and G15 on the basis of the observed number of new alleles appearing between the two generations studied. The gene-
genealogy estimators and their derivation are described in the appendix. The results about the difference between the estimators are detailed and discussed in the appendix. In short they provided fairly consistent estimates (Table 2). Whenever computable, the $\mu_1$ and $\mu_2$ estimators based on the continuous time Kingman model provided similar results, suggesting that the correlation among internode durations was not a major effect for our set of parameter values. These two estimators gave values close to that of $\mu_3$ and the simulated $\mu_4$ (Wright-Fisher model), but provided a slightly greater estimate of the mutation rate. The first two estimators probably underestimated the size of the tree, mainly because of the continuous time approximation: with our set of parameter values, many coalescent events would be estimated to occur before a single generation ended. Multiple common ancestries occurring at the same generation would introduce a bias in the opposite direction and seemed therefore to affect the estimates to a lesser extent. The $\mu_3$ and $\mu_4$ estimators based on the same - more appropriate - WF model provided virtually identical results, showing that the (A4) analytic approximation is a reliable approximation. We only report below results from the estimator $\mu_4$ based on genealogy simulations, the closest model to our experimental design, which should provide the most reliable estimates and allows for computation of standard errors, though other estimators may well be more useful for other set of parameter values and design.

We tested the distribution homogeneity of the number of new alleles across the whole set of 21 loci between parents and G15 (see appendix). The distribution of new alleles across loci was the following: (0, 1, 2, 3, 4) new alleles at (12, 2, 3, 1, 3) loci respectively. The parameter of the Poisson distribution was estimated as 1.095. Because of the restricted number of loci we split the distribution into three classes (no new alleles, one new allele and two or more new alleles). Due to the 0.005 probability associated with the Chi square test we rejected the null hypothesis of homogeneity of the number of new alleles across the 21 loci. The comparison between the two distributions revealed an excess of loci with no new alleles and a
deficit of loci with one new allele. Loci with new alleles did not show any clear discrepancy with a Poisson distribution which could lead to different realized genealogy size. Thus, we considered the nine loci with new alleles as homogeneous.

For these nine loci, the mean mutation rate estimated between the parents and the 15th generation was 1.68 x 10^{-3}. The 95% confidence interval was estimated as [1.50 x10^{-3}; 1.86 x10^{-3}]. The highest estimates of the mutation rate were obtained between parents and G1 (calculated on six loci), where $\mu_4$ ranged from 1.29 x 10^{-3} to 4.79 x 10^{-3} with an average of 2.98 x 10^{-3} (CI_{95%} = [2.71 x 10^{-3}; 3.26 x 10^{-3}]). The averaged estimation for G1-G15 was 0.93 x 10^{-3} (CI_{95%} = [0.79 x 10^{-3}; 1.07 x 10^{-3}]). Note that the six mutation rates between G1 and G15 all had smaller estimated values than their counterparts calculated between parents and G1. Under the assumptions of a constant mutation rate whatever the period (Par-G1 or G1-G15) and no bias of estimation linked with the period, the expected probability of this event (six estimates out of six with a lower value in the second phase) was 0.0156 (1/2^6). So we can reject the null hypothesis at the 5% level.

**Factors influencing locus diversity and mutation:** A significant effect of the number of repeats in the microsatellite motif was found on the number of new alleles detected in G15 (P<0.05, $R^2 = 0.21$), but not on the number of parental alleles. The Nei polymorphism index ($H$) calculated on the French wheat collection with the same 21 loci had a highly significant correlation with the number of parental alleles (P<0.01, $R^2 = 0.54$) and with the number of new alleles observed in the population (P<0.01, $R^2 = 0.33$). Similarly, the number of alleles detected with the same set of markers on the wheat collection was correlated with the number of parental (P<0.01, $R^2 = 0.52$) and new alleles (P<0.01, $R^2 = 0.40$).
DISCUSSION

Mutation rate estimators

We used an experimental wheat population to assess mutation rates at microsatellite loci using four new gene genealogy based methods. We found two classes of loci: 12 without mutation, nine with mutation rates estimated as $10^{-3}$-$10^{-4}$ which were positively correlated to the length of the microsatellites.

The methods developed here based on gene genealogy approaches proved to be useful to estimate mutation rates on population with temporally spaced samples. The present experimental design facilitates such estimation since the initial variation is entirely known and all newly arisen mutations can be unambiguously identified. Thus it is only necessary to estimate the size of the tree. This might well not be a common case. It could, however, also apply to experiments starting from a clonal monomorphic population such as those involving microorganisms. A more frequent problem concerns time series data, e.g. ancient DNA data, pathogen evolution and other empirical studies with several sampling points in time, where the initial variation is not exhaustively typed. Extension of the method would then require adding explicitly a mutational model and more elaborate method such as MCMC Bayesian or frequentist approaches (Drummond et al. 2002). Such a Bayesian approach can be adapted to our model (Kingman continuous time approximation; M. Blum personal communication). It can be extended to allow for a more complex mutational model taking into account the complicating homoplasy issue common with microsatellite data. A classical (power) issue is to disentangle the mutation rate from the effective size. Independently obtaining an estimate of the effective size, as in the present case, whenever possible, would certainly help. Moreover, it is not clear that there would be much power to assess which part of the newly detected variation is due to mutations or to ancestral variation not sampled before. An estimate of the mutation rate could combine both time scale information, variability pre-
existing the sampling and newly arisen variability. Which one would be most useful to estimate the mutation rate remains unclear, being highly dependent on the sampling scheme and set of parameter values (see the appendix for a discussion of other related issues on gene genealogy based methods).

**Mutation vs migration**

Although the experimental population was grown in isolation from other wheat cultures, the occurrence of new alleles at subsequent generations can be either due to migration or to mutation. Here, independent observations and analyses provided evidence that most new alleles were generated by mutation:

- The emergence of new alleles at different loci between the 16 parents of the population and the generations studied (G1, G5, G10, G15) suggested that the new alleles appeared independently and therefore, there was no indication that they could be due to recent migration events.

- Moreover, their presence could not be explained by migration of pollen or seeds of the wheat varieties identified as possible sources of migration because the new alleles did not correspond to the alleles of these varieties, or when they did correspond, their respective growing years were not compatible.

- Sequencing results showed polymorphism in the core of the microsatellite, for instance, the three alleles cloned at locus *Xgwm642* only differed by their number of repeats, and no indel in the flanking sequences was observed. Most new alleles differed from parental alleles by only one or a few microsatellite repeats in agreement with stepwise like mutational models.

**Comparison with previous mutation rate estimates**
The estimates of mutation rates obtained in this study for loci showing new mutations ($10^{-3}$-$10^{-4}$) were consistent with but rather in the upper range of mutation rates estimated at microsatellite loci in *Triticum turgidum* (Thuillet et al. 2002) and *Zea mays* (Vigouroux et al. 2002) based on mutation accumulation lines, and also on distantly related species like barn swallows (Brohede et al. 2004), avian (Beck et al. 2003), *Gastropoda* (Gow et al. 2005) and humans (Weber and Wong 1993; Heyer et al. 1997; Whittaker 2003) with pedigree or parent-offspring based estimations. In contrast, estimations in mutations accumulation lines of drosophila were smaller ($\sim 10^{-6}$, Schug et al. 1998; Vasquez et al. 2000) possibly due to shorter microsatellite sequences, as were estimations based on inter-specific comparison between humans and chimps ($\sim 10^{-4}$-$10^{-5}$, Webster et al. 2002; Sainudiin 2004) which might suffer from the confounding effects of selection, demography and risks of saturation.

In natural populations, microsatellite mutation rates might be higher than in highly homozygous accumulation lines because the high number of heterozygotes will increase microsatellite instability by increasing the probability of unequal crossing-over (Rubinsztein et al. 1995). Consistently, the number of new alleles already detected in G1 (which was separated from the parents by four generations of inter-crosses followed by four generations of multiplications) was higher than the number found over the 15 generations separating G1 from G15. This should not be due to differences in $N_e g$ since the estimated $N_e g$ over different periods, whether including the inter-crosses and multiplication or not, were very similar, as previously detailed. Homoplasy should mostly lead to substantial underestimation of mutation rates in natural populations compared to studies using accumulation lines. Here we provide the first study reporting reliable estimates of mutation rates in an evolving population.

**Mutation model**

The observed distribution of the new alleles allowed us to reject the null hypothesis of homogeneity among the 21 loci with an excess of loci with no new alleles. In contrast, no
heterogeneity could be detected among the nine loci with new alleles. Heterogeneity may be due either to different mutation rates or different realized genealogy sizes. If it is due to different tree sizes, this could reflect either different effective sizes and/or variance in tree size for a given effective size. Based on their variations in allelic frequencies, two loci were detected as outliers with locally reduced genetic effective population size. Given that they were among the nine with new alleles, these \( N_{ee} \) differences were unlikely to cause substantial heterogeneity in tree sizes among the remaining 19 (\( N_e \)) homogeneous loci. For a given (estimated) effective size, our simulations revealed a low (4%) coefficient of variation of partial tree sizes (\( L_T \)). This is largely due to the fact that the portion of the tree involved during the experiment corresponds to the bottom part of a neutral constant size coalescent tree, averaging across many coalescence events (thus reducing variances). In contrast, the top part of a coalescent tree –not reproduced during our experiment- usually relies on very few highly stochastic coalescence events and this leads to a large stochastic variance in the age of the MRCA and thus, in the tree size. Therefore, in the present study, we can speculate that the heterogeneity was mainly due to differences in mutation rates. To get deeper insight into the mechanisms generating diversity, we studied the effects of some microsatellite intrinsic factors on the number of new alleles in G15. Among the different factors, the motif repeat number of the microsatellite was found to be significantly correlated to the number of new alleles in the G15 generation. This result is consistent with mutation studies that suggest that longer microsatellites are more unstable (Wierdl et al. 1997) and have higher mutation rates (Rubinsztein et al. 1995; Schug et al. 1998; Seyfert et al. 2008). In this study, we found no evidence that the motif of the microsatellite locus, its degree of perfection, its chromosomal location or its genome (A, B, D) position had an effect on the microsatellite diversity, but we had little power to detect weak effects. The literature is inconsistent on these topics on wheat
(Maccaferri et al. 2003; Thuillet et al. 2004) or on more distant species like humans (Boyer et al., 2008).

Comparison between allele sequences of bread wheat, durum wheat (A and B genomes) and Aegilops tauschii (bread wheat wild ancestor with D genome) showed a fairly good conservation of flanking sequences within bread wheat and between the three species, indicating that polymorphism at these microsatellite loci was due to differences in the number of repeats. Among the nine loci studied, five had new alleles differing from parental alleles by only one microsatellite repeat. This result supported the Stepwise Mutation Model which assumes that new alleles are created by an increase or a decrease of one repeat (Kimura and Ohta 1978). For the four other loci, results were more consistent with the Two-Phase Model which assumes that some mutations are created by addition or deletion of one repeat, while others are created by the indels of several repeats (Di Rienzo et al. 1994). Among the 26 new alleles observed in the population, seven alleles differed from parental alleles by more than one repeat. While this novel diversity might have arisen from unequal crossing-over during recombination (Sia et al. 1997), the 19 other new alleles could have arisen either from unequal crossing-over or from polymerase slippage during replication. It was not possible to assess the relative contributions of these two mechanisms.

Mutation has a positive impact on genetic diversity (whose maximal value at a given locus is determined by the number of alleles at this locus) and obviously on allelic richness, as exemplified in our experimental study: out of 86 alleles initially present, only three were not detected in the G15 generation. Moreover, 23 new alleles appeared in the population, resulting in a gain of 20 alleles. Nei’s diversity was lower in G15 than in the parents for four loci, but was otherwise similar between these two generations.
On the other hand, it is not likely that mutation influences the evolution of allele frequencies over time except in case of recurrent mutation over a large number of generations. Rather, temporal variation in allele frequencies reflects pressures such as genetic drift and selection.

**Evolutionary regime**

Here, the estimated effective size based on allele frequency variation between parents and the generation G15 ($N_{ef}$=183) was much smaller than the demographic effective size ($N_{ed}$=2625) based on the true number of plants cultivated at each generation. The difference could not be explained by genetic drift alone but selection was suspected.

It is normal for $N_{ef}$ to be lower than $N_{ed}$, as shown in a review of 192 experiments by Frankham (1995). He analyzed the $N_e/N$ ratios measured in natural animal and plant populations and the mean ratio was about 0.10, with the most important variables influencing the value of this ratio being the fluctuation in population size and variance in family size. Yet, in the experimental wheat population, the number of individuals was controlled, so that very few demographic fluctuations might have happened. Here, the low $N_{ef}$ ($=183$) indicates that all plants did not contribute equally to subsequent generations. Instead, there was a large variance in reproductive contribution. This variance in reproductive contribution may be due either to non-inherited causes or to inherited causes. In the latter case, selection would increase the variance of reproductive contribution (indirect effect) while modifying the frequency of the genes involved in the control of the fitness-related traits. In a previous study (Goldringer et al. 2001), using the analytical formulae (Caballero 1994) for the calculation of $N_{ef}$ in the case of non-inherited variation in the reproductive contribution, we found that only unrealistic variances (much greater than direct empirical estimates) could explain the discrepancy observed between the estimated $N_{ef}$ and the demographic size of the populations. On the other hand, theory (Santiago and Caballero, 1995) and some experimental results (Austerlitz and Heyer, 1998) showed that low levels of correlation among the effective family
sizes for successive generations may strongly reduce genetic effective population size in a quantitative selection way. This led us to conclude that differences in parental contributions were due to inherited genetic causes and that, in other terms, limited effective size in the population was due to selection. The two loci (Xgwm312 and Xgwm642) with the highest (but not significant) individual $F_{c}$-values also presented new alleles. Introduction of new alleles by mutation is unlikely to drive high $F_{c}$-values. For the locus Xgwm312, new alleles were detected at low frequencies (<5%) at each generation and high $F_{c}$-value was only explained by strong variation in parental allele frequencies. For the locus Xgwm642, the new allele frequency increased across generations, and the high $F_{c}$-value estimated was due to strong variation of parental and new allele frequencies. This is likely to be due to the selection of favorable variants at closely linked loci (hitch-hiking effect, Maynard Smith and Haigh, 1974).

The effective size calculated here was also similar to those calculated in previous studies using RFLP markers in the first ten generations (Goldringer et al. 2001). Despite large differences in the number of alleles detected per locus in the population and in the mutational process between RFLP markers (average of 2.2 alleles per locus in G10, Goldringer et al., 2001) and microsatellite markers (average of 5.0 alleles per locus in G15), using both kinds of markers led to very similar estimated effective size.

Selection pressures assumed to reinforce genetic drift effects on allele frequencies did not decrease the initial variability for most of the loci studied here. While drift and directional selection should increase the probability of allele fixation, mutation and migration allow for the renewal of the allelic stock. Here, allelic richness and genetic diversity already existing for microsatellite markers in the parental population was mostly conserved after 22 generations of evolution in dynamic management, suggesting that for microsatellite markers, emergence of diversity by mutation balanced the lost by selection and drift. Studying allelic diversity and its
qualitative and quantitative variations over time allowed us to identify the significant evolutionary forces (mutation and selection) working on this population and their effects on the conservation of neutral diversity.
Acknowledgments

We thank C. Baron and P. Sourdille for the genotyping of the G15 generation, P. Leroy for providing the public ITMI map data and S. Santoni for allele sequencing. We are grateful to T. Bataillon, D. Lachaise and R. Bernardo for their useful comments. This work was supported by a grant from the Bureau des Ressources Génétiques (BRG) and ACI IMPBIO (French Ministry of Research) to FD and AL.
Literature Cited


Kayser, M., E.J. Vowles, D. Kappei and W. Amos, 2006 Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal loci. Genetics 173: 2179-2186.


Nei, M., and F. Tajima, 1981 Genetic drift and estimation of effective population size. 
Genetics **98**: 625-640.

Pejic, I., P. Ajmone-Marsan, M. Morgante, V. Kozumplick, P. Castiglioni et al., 1998 
Comparative analysis of genetic analysis among maize inbred lines detected by 
RFLPs, RAPDs, SSRs, and AFLPs. Theor. Appl. Genet. **91**: 1001-1007.

Plaschke, J., M. W. Ganal and M. S. Roder, 1995 Detection of Genetic Diversity in Closely-
Related Bread Wheat Using Microsatellite Markers. Theoretical and Applied Genetics 
**91**: 1001-1007.

Röder, M. S., V. Korzun, K. Wendehake, J. Plaschke, M. H. Tixier et al., 1998 A 

Roussel, V., J. Koenig, M. Beckert and F. Balfourier, 2004 Molecular diversity in French 
bread wheat accessions related to temporal trends and breeding programmes. Theor 
Appl Genet **108**: 920-930.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain et al., 1995 Microsatellite 
evolution--evidence for directionality and variation in rate between species. Nat Genet 
**10**: 337-343.

Sainudiin R., R. T. Durrett, C. F. Aquadro and R. Nielsen, 2004 Microsatellite mutation 
models: insights from a comparison of humans and chimpanzees. Genetics **168**: 383-
395.

Santiago E., and A. Caballero, 1995 Effective size of populations under selection. Genetics 
**139** (2): 1013-1030.


Schlotterer, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109**: 365-
371.


Appendices: Derivation of gene genealogy estimators of the mutation rate $\mu$

We consider the issue in terms of gene genealogies (Kingman 1982), as a general and intuitive framework to make inferences in population genetics.

**Main assumptions:**

*Demography*

In general it is difficult to disentangle the mutation rate from the effective size and a compound parameter is commonly inferred: the mutational parameter of the population $\theta=4N_e\mu$ (for a diploid population, autosomal loci). For the present data set, as in some other instances, we have an independent way of estimating the effective size. Thus it is straightforward to deduce an estimate of the mutation rate from estimates of $\theta$. Indeed, temporal samples are available, so it is possible to estimate the variance effective population size of the population, through temporal variations of the frequency of the ancestral alleles (preexisting the experiment; see above, diversity sections of the methods and results). This summary parameter provides a way to capture the demographic effect, which we shall neglect (non panmictic population with variable size). In our experimental application case, this should typically concern the specific crossing-regime and increase of (seed) population size at the beginning stage of the experiment. Estimates were fairly consistent between the different time steps of the experiment, suggesting robustness of the estimators to such effects (see application below).

*Mutations*

We assume that there is no homoplasy, that is, all new $S_l$ alleles detected in the sample for locus $l$ have a unique mutational origin. However, it is possible that a neomutant arising from one of the original variety’s descendents gives rise to an allele identical to that of another original variety. Then the results would depend on the detailed configuration of the
parental sample. But many of those parental variety alleles differ substantially in length size. Such homoplasies become rather unlikely under a stepwise like model. Remaining cases should lead to a slight underestimate of the mutation rate. In addition, the assumptions are similar to the infinitely many sites model (ISM; Watterson 1975), in the sense that we assume that all mutations occurring in the history of the sample since the beginning of the experiment are detected in the present sample. Note that our model is even more stringent than the infinite allele model (IAM). Roughly speaking, it is assumed that each mutant occurring in the genealogy is represented in the resulting sample (i.e. that it gives rise to at least one descendant). This seems nevertheless realistic given the small number of newly arisen mutations detected in the samples and the large number of branches in the genealogy (there remain typically roughly 30 distinct lineages at the beginning of the experiment; see below). Indeed, a sufficient condition for all newly arisen mutants to be detected is that they appear on different subtrees descending from the various ancestors. In the worst case scenario, with four mutations detected on a locus, the probability that a fifth one is undetected is thus below $4/30$ (13%) and the probability for a sixth one to remain undetected is $(4/30)^2$ (1.7%). These probabilities clearly indicate that we are unlikely to grossly underestimate the mutation rate. Nevertheless, for two of our three analytical approaches $\mu_1$ and $\mu_3$ see below), we extended the results to an infinitely many alleles model (IAM), with virtually no change on the results (slightly increased estimates of $\mu$ as expected, see below).

**Preliminaries**

Since mutations are supposed to occur at a constant rate $\mu$, the total number of mutations occurring during the genealogy of a standing sample of size $n$ is a Poisson variable with parameter and expectation $\mu L$, where $L$ is the total length of the genealogy also commonly called the size of the tree. In population genetics terms, $L$ is the sum of all life spans of the common ancestors during the genealogy of the sample, that is, the real time duration during
which currently observable mutations may have occurred. Under an ISM-type model, where
all mutations occurring on the tree are readily detectable in the resulting sample, a simple
estimator of the mutation parameter is provided by dividing \( S \) the observed number of
mutations in the current sample by \( L \) (Watterson 1975).

*Partial tree*

However, in our experimental application, the initial (ancestral) state of the population is
entirely known and the process needs to be considered only for the duration \( T \) of the
experiment and typically not for the whole duration of the genealogy of the sample. The rest
of the genealogy is thus considered as a black box leading to the observed initial point of the
experiment after which (going forward in time) conditions and parameters are most controlled
and finely monitored, which should limit drastically putative biases.

In particular, \( S_T \) the number of new mutations having occurred *since the beginning* of the
experiment is exactly known (under the above mentioned mutational assumptions) and our
mean estimate of the mutation rate is then provided by dividing this number by the size of the
*partial* genealogy \( L_T \) since the beginning of the experiment.

We denote by \( X_t \) the number of ancestors of the sample \( t \) units of time ago \( (X_0 = n) \). As we
shall see, it is not straightforward to extend general coalescent results to the partial tree case
since corresponding properties should depend on the initial state, *i.e.* on a random variable the
number of distinct ancestral founders of the sample \( T \) units of time ago \( X_T = p_\).  

The goal of this appendix is thus to investigate the ancestral state at this given time point, the
number of ancestors \( X_T \) and above all the corresponding *size* \( L_T \) of the partial tree.

**Size of partial coalescent trees**

We propose four different methods for estimating the size of the partial tree leading to four
different estimators of the mutation rate, \( \mu_1, \mu_2, \mu_3 \) and \( \mu_4 \) respectively. They rely on two
different models of gene genealogies, depending on how time is modeled. The first two estimators are analytical estimators obtained under a continuous time model (I), classically referred to as Kingman’s (1982) coalescent process. The $\mu_1$ estimator neglects internode correlation induced by the partial tree conditioning (see below), while $\mu_2$ stems from an exact analytical derivation. The latter two estimators are obtained under and a discrete time analogue, the Wright-Fisher model (II). The $\mu_3$ estimator is an analytical approximation computed recursively and $\mu_4$ is based on genealogy simulations. Times and in particular $T$ are thus a real numbers for model I, and integers (number of generations) for model II.

1 Model I Kingman’s coalescent (continuous time model)
In this model, pairs of genes coalesce at constant rate $c=1/2N_e$ per time unit (generation). In other words, the most closely related pair of genes, in a sample of $k$ genes, has a closest common ancestor that lived some random time $\tau_k$ ago, where $\tau_k$ is an exponential variable with parameter $c_k=c k (k - 1)/2$.

1.1 $\mu_1$: neglecting internode correlation
In Kingman’s coalescent, it is easy to compute the size of the tree $L_k$, not for a fixed time $T$, but for a given number $n-k+1$ of common ancestry events. Indeed, let $S_n=0$ and $S_k$ be the time of the $n-k+1$-th coalescence event backwards in time, so that

$$S_k \leq t < S_{k-1} \iff X_t = k,$$

meaning that the number of ancestor genes is $k$ between times $S_k$ and $S_{k-1}$.

Then observe that the total length of the tree since time $S_p$ is equal to

$$L_p = \sum_{k=p+1}^{n} k \tau_k.$$
where $\tau_k = S_k - 1$ is the duration of internode $k$. However, when $T$ is a fixed time, it depends on the number of ancestors at the beginning of the experiment $X_T = p$. Conditional on a given number $p$ of ancestors, $S_p$ is the last coalescence time, and

$$L_T = p(T - S_p) + \sum_{k=p+1}^{n} k \tau_k.$$  

(A1)

As in the standard full genealogy case, the computation of expectation and variance of the tree length are simplified by the independence of the $\tau_k$’s. The expectation of the tree size is then simply computed from (A1), by replacing the internode durations $\tau_k$ by their expectations $1/c_k$ under the Kingman (1982) coalescent and computing the summation (A1) as long as the sum of the $\tau_k$’s $(S_{k,1})$ is not greater than $T$ (continuous time model). Then $p$ ancestral lineages remain and the residual term of (A1) $p(T - S_p)$ is added.

But strictly speaking, when conditioning the model on an absolute duration $T$, the internode durations become non independent from one another since they are conditioned upon their sum $S_p$ remaining lower than $T$ and analytical results are therefore more difficult to obtain.

1.2 $\mu_2$: taking into account internode correlations, in Kingman’s coalescent, exact formulae for the expected tree length and the expected number of alleles.

We can show by induction that the Laplace transform of $L_T$ on the event that there are $p$ ancestor genes in the population backward to time $T$ equals:

$$E(e^{-\mu_T}, X_T = p) = \left( \prod_{i=p+1}^{n} c_i \right) \sum_{k=p}^{n} \frac{e^{-T(c_i + k\lambda)}}{\prod_{i=p,j=k}^{n} (c_i - c_k - i\lambda - k\lambda)}$$

Taking this formula at $\lambda = 0$, as well as its derivative at $\lambda = 0$, yields respectively the probability that $X_T = p$ (the distribution of the number of ancestor time $T$ ago), and the expectation of $L_T$ with $X_T = p$: 


\begin{equation}
P(X_T = p) = \left( \prod_{i=p+1}^{n} c_i \right) \sum_{k=p}^{n} \frac{e^{-Tc}}{\prod_{i=p,i \neq k}^{n} (c_i - c_k)} \end{equation}

\begin{equation}
E(L_T, X_T = p) = \left( \prod_{i=p+1}^{n} c_i \right) \sum_{k=p}^{n} \frac{e^{-Tc}}{\prod_{i=p,i \neq k}^{n} (c_i - c_k)} \left[ kT + \sum_{i=p,i \neq k}^{n} \frac{2}{c(i + k - 1)} \right] \tag{A2}
\end{equation}

It is then straightforward to get the expectation of $L_T$, since

\begin{equation}
E(L_T) = \sum_{p=1}^{n} E(L_T, X_T = p) \tag{A3}
\end{equation}

However, in this method, a numerical drift issue arose in the computation of the numerical results for some extreme sets of parameter values (low $T$, high $n$ with respect to $N_e$). The above first two approaches are in line with the standard Kingman (1982) coalescent, assuming a binary tree (with no higher order multifurcation, e.g. three individuals showing the same common ancestor or several pairs of individuals reaching their common ancestors at the same generation). They also assume a continuous time approximation, replacing the geometric distribution for the coalescence times by its exponential limit. Both approximations are intended to apply to populations with large effective size with respect to the sample size and long coalescence time in generation units. Because of this limitation, this approach may not be appropriate for our (and many other) set(s) of parameter values with a small number of generations surveyed and large sample size and with respect to the effective size.

To overcome this weakness, we now turn to a classical Wright-Fisher model

**Model 2: Wright-Fisher**

The most prominent differences with the previous model are that time is expressed in discrete generations, and that more than two genes are allowed to coalesce at the same time (generation).
Each individual gene has exactly one unit of time to undergo mutation before it is passed to the next generation, so that $L_T$ is simply the sum of the $X_t$'s for $t = 0,\ldots,T$.

2.1 $\mu_3$: a simple approximation formula

It is easy to see that $X_0, X_1, X_2,\ldots$ is a homogeneous non-increasing Markov chain stopped at 1. In addition, let $p_k^{(i)}$ be the probability for $k$ genes to have $i$ ancestor genes at the previous generation

$$p_k^{(i)} = P(X_{t+1} = i|X_t = k).$$

It is then possible to get a recurrence relationship for these transition probabilities. Namely, consider a sample of $k+1$ genes as a sample of $k$ genes plus one extra gene. This ($k+1$)-sample has $i$ ancestor genes at the previous generation if either the $k$-sample has $i$ ancestors which include the extra gene's ancestor, or the $k$-sample has $i-1$ ancestors and the extra gene has a distinct ancestor. This can be written in the neutral setting as

$$p_{k+1}^{(i)} = \left(1 - \frac{i-1}{2N_e}\right)p_k^{(i-1)} + \frac{i}{2N_e}p_k^{(i)},$$

for any $i \leq \min(2N_e,k)$. Then let $f_k$ be the generating function of $X_t$ when $X_{t-1}=k$, and $\gamma = 1/2N_e$.

The last recurrence relationship then translates into

$$f_{k+1}(s) = sf_k(s) + \gamma s(1-s)f_k'(s),$$

with the condition that $f_1(s) = s$ for any $s$ in the interval $[0,1]$. Taking derivatives at $s=1$ yields the following equation for the expectation of $X_t$:

$$E(X_t | X_{t-1} = k) = g(k),$$

where

$$g(x) = \frac{1}{\gamma} \left(1 - e^{-\alpha x}\right),$$

42
and $\alpha = -\ln(1 - \gamma)$. Then a simple approximation for the expectation of $X_2$ is $Y_2 = g(g(n))$, which is equivalent to replacing $X_1$ by its expectation in the $g$ function. More generally we take the $t$-fold composition $g'$ of $g$ for the approximation $Y_t$ of $E(X_t)$ (Fig. A2).

To conclude, the total length of the tree is exactly the sum of the $X_i$s, so the total length of the tree during the last $T$ generations can be approximated by

$$L_T = \sum_{t=0}^{T-1} g'(n) = \sum_{t=0}^{T-1} Y_t$$

(A6)

To check the limit of this approximation by recursively replacing the number of ancestors by their conditional expectation, we used a simulation approach.

2.2) $\mu_4$: Monte-Carlo simulation of the WF model

We used a backward coalescent-related Monte Carlo simulation algorithm to estimate the mean size of the trees. The principle of the algorithm is as follows: it proceeds backwards, generation by generation. It starts with $X_0 = n$ genotyped individuals in the sample, with the size of the tree $L_0$ initialized to 0. To account for the first generation, $X_0$ is added to the current size of the tree ($X_0$ meioses implicitly occur across the current generation in the history of the sample). To get the previous generation, for each of the $X_0$ individuals in the sample, an ancestor is randomly drawn (with replacement) among the $N_e$ possible ones available in the previous generation. We then compute $X_1$ the resulting number of distinct ancestors drawn as ancestral to the sample. To move to the previous generation, $X_1$ replaces $X_0$. The process is repeated for the $T$ generations of the experiment.

Variance of the mutation rate estimate

We assume that the number $S$ of mutations is Poisson distributed with parameter $\mu L_T$, where $\mu$ is the mutation rate, independent from the total length $L_T$ of the tree. The estimator of $\mu$ is simply $\hat{\mu} = S / L_T$. It is then well-known that $\hat{\mu}$ is unbiased since

$$E(\hat{\mu}) = E(E(\hat{\mu}|L_T)) = E(E(S|L_T)|L_T)) = E(E(S|L_T)/L_T) = E(\mu L_T / L_T) = \mu.$$
The same computation can be done for the variance of \( \hat{\mu} \)

\[
V(\hat{\mu}) = E(V(\hat{\mu}|L_T)) = E(V(S/L_T | L_T)) = E(V(S|L_T) / L_T^2) = E(\mu L_T / L_T^2) = \mu E(1/L_T),
\]

so that the variance of \( \hat{\mu} \) is equal to the mutation rate divided by the harmonic mean of \( L_T \).

Hence, an estimate of this standard variance is obtained by dividing the estimate \( \mu \hat{\lambda} \) by the harmonic mean of \( L_T \) across the simulations, which is easily obtained in this approach since it empirically provides the full distribution of the tree size.

At this point of this section, we want to point out that the data we have are given in terms of numbers of alleles, and not numbers of mutations. If there were no intermediate (not detected in the final sample) mutations, these two numbers would be equal. To take into account that it might not be the case, it is standard to consider an IAM-type model.

**Infinite allele case**

In both ISM and IAM, it is possible to derive the expected number of mutations or alleles (respectively) as a function of the mutation rate. In ISM this function is linear, so that it is straightforward to deduce an estimator of the mutation rate from the expected number of mutations. This is not the case in the IAM, although it is in principle possible to numerically invert this function to estimate the mutation rate given the observed number of alleles \( K_T \).

A simple argument allows however for a straightforward derivation of an estimate. Under IAM, while tracing the lineages backwards in time, both common ancestry and/or mutation events can occur. Each time a mutation is encountered before time \( T \), a new allele is added to the sample and the corresponding lineage is stopped, since further mutations will not increment nor decrement the number of alleles \( K_T \) in the final sample (Griffiths, 1980; Tavaré 1984). Then the expected value of \( K_T \) is calculated by adding the probability of mutation to that of common ancestry in the exponential parameters of inter-event durations, and the
number of alleles in the sample is simply the number of mutation events having occurred
before time $T$, which is given by

$$E(K_T) = \sum_{i=2}^{n} \frac{\mu}{\mu + c(i-1)/2} P(\tau_i < T) \quad (A4)$$

In the approximation case $\mu_1$, the rightmost term takes only marginal values 0 or 1, and (A4)
can then be easily computed and inverted numerically for $\mu$.

In the more exact correlated internode case (above section 1.2), the last term of (A4) can be
computed recursively as

$$P(\tau_i < T) = 1 - \sum_{j=1}^{n} \prod_{k=1,k\neq j}^{n} \left( \frac{e^{-(c_j+\mu j)t}}{e^{c_j+\mu k}-1} \right)$$

But inverting it numerically for $\mu$ turns out to be cumbersome, due to the above mentioned
numerical drift issue and we do not present corresponding results.

In the Wright Fisher approximation case ($\mu_3$), we define $Z_t$ similarly as the number of ancestor
lineages in generation $t$ backwards in time that have not experienced any mutations so far ($Z_0$
equals $n$). Then $Z_t$ is computed as the $t$-fold composition of $h(x)=g((1-\mu)x)$, and $\mu Z_t$ accounts
for the expected number of new mutations having occurred in generation $y$, which can then be
summed over the $T$ generations of the experiment to approximate the expected $K_T$. This can
then be solved numerically for $\mu$.

In the simulation approach $\mu_4$, it is easy to compute the mean $K_T$ given the mutation rate by
drawing in each generation the number of mutant offspring according to a binomial and then
removing them from further ancestry. However, estimating the mutation rate from this
approach would require more computationally intensive simulation schemes, for example
with Bayesian approaches, which seems beyond the scope of the paper given the weakness of
the effect involved (see below).

Case study:
Mutation rates were estimated for each of the six loci studied at intermediate generations between parents and G1, and between G1 and G15, and were estimated for each of the nine loci with new alleles between parents and G15. To estimate the mutation rates, we used the $N_e$ estimated between parents and the 15th generation on the 21 loci. Under the null hypothesis “all loci follow the same distribution as to their coalescence time and mutation processes (and in particular show the same mutation rate and $N_e$),” we compared the observed distribution of the number of new alleles to the expected null Poisson distribution of parameter $\mu L_T$ using a Chi-square test, where $\mu$ is the mean estimate over loci and $L_T$ the expected size of the partial tree. Under these assumptions the Poisson parameter was simply estimated as the mean number of new alleles per locus. The test gave a rough approximation of the homogeneous or heterogeneous behavior of the different loci with respect to their mutation rate and coalescence time. Indeed, rejection of the null hypothesis would only mean that $\mu L_T$ is not constant across loci with no possibility for separating the two possible causes: different mutation rates across loci or different realized $L_T$ (either derived from the same $N_e$ value and due to stochasticity in the coalescent or due to different $N_e$ values). For the loci considered to have a homogeneous distribution of new alleles, we applied the central limit theorem to derive the confidence interval of the mean mutation rate across loci $[\mu - 1.96 \cdot \text{SE}; \mu + 1.96 \cdot \text{SE}]$.

The mutational model showed little effect on the estimates, providing slightly greater estimates with IAM as expected (the largest difference was $0.12 \times 10^{-3}$ for $\mu_1$ at locus Cfd71_D between parents and G15, results not shown). Similarly, we expect that homoplasy, due to recurrent or reverse mutation occurring during the partial genealogy in any finite allele or stepwise like mutational models, should not affect much the estimates.

The $\mu_1$ and $\mu_2$ estimators based on the continuous time Kingman model provided similar results, suggesting that the correlation among internode durations was not a major effect for our set of parameter values (Table 2). The effect should however be greater if the number of
ancestors p at the beginning of the process was smaller. Due to a numerical drift issue for a small number of generations, μ2 could not be estimated between parents and G1, and between G1 and G15 in two cases.

For larger time steps, both μ1 and μ2 were close to the simulated μ4 (Wright-Fisher model), but provided a slightly greater estimate of the mutation rate (Table 2). These estimators probably underestimated the size of the tree and it appeared useful to relax some of the main assumptions of the standard coalescent. Two factors seemed to act in a synergetic way in this study, where many common ancestries could occur before the first generation in the standard coalescent setting. First, it is probably most relevant to allow for multiple common ancestries in such empirical cases, where the sample size might be large compared to the effective size, partly because the census size is much greater than the effective size. This might occur for instance in conservation related issues though in our case even if most common ancestry event occur in the first few generations they contribute substantially to the partial tree size given the limited total number of generations of the experiment. Since the coalescent process in continuous time does not allow for multifurcation, it may introduce a bias in the opposite direction, that is, it should put more weight on larger trees. Our results probably reflected a bias due to the second factor: the continuous time approximation. Under this assumption, many lineages did indeed coalesce before the end of the first generation.

A striking result is that μ3 and μ4 estimators based on the same Wright Fisher model provided virtually identical results (Table 2), showing that our approximation formula (A4) is a quick but accurate estimate of the tree size in the WF model, with very little numerical drift.

As to the effect of putative selection on our estimators, we think the method should be quite robust in comparison to such effects since it is not influenced by the frequency of newly arisen mutations, which could be shifted by selection. Finally, the method relies on an estimate of the effective size that takes into account the overall effect of selection.
A Mathematica notebook implementing the above estimators is available upon request to Frantz Depaulis.
Table 1: frequency trajectories and emergence of new alleles in the population.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele size (in bp)</th>
<th>Allele frequency</th>
<th>Status</th>
<th>Difference between number of motifs compared with parental alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Locus</td>
<td>Parents</td>
<td>PA1</td>
<td>PA5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fraction number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>141</td>
<td>4/16</td>
<td>0.250</td>
<td>0.278</td>
<td>0.232</td>
</tr>
<tr>
<td>143</td>
<td>1/16</td>
<td>0.063</td>
<td>0.062</td>
<td>0.133</td>
</tr>
<tr>
<td>145</td>
<td>-</td>
<td>-</td>
<td>0.003</td>
<td>-</td>
</tr>
<tr>
<td>147</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>151</td>
<td>8/16</td>
<td>0.500</td>
<td>0.373</td>
<td>0.377</td>
</tr>
<tr>
<td>Xgwm181</td>
<td>153</td>
<td>-</td>
<td>-</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>2/16</td>
<td>0.125</td>
<td>0.219</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>1/16</td>
<td>0.063</td>
<td>0.029</td>
</tr>
<tr>
<td>162</td>
<td>-</td>
<td>-</td>
<td>0.069</td>
<td>0.075</td>
</tr>
<tr>
<td>174</td>
<td>1/16</td>
<td>0.063</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>176</td>
<td>1/16</td>
<td>0.063</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td>6/16</td>
<td>0.375</td>
<td>0.392</td>
</tr>
<tr>
<td>Xgwm260</td>
<td>180</td>
<td>7/16</td>
<td>0.438</td>
<td>0.379</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td>-</td>
<td>-</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>184</td>
<td>-</td>
<td>-</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>186</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>1/16</td>
<td>0.063</td>
<td>0.086</td>
</tr>
<tr>
<td>130</td>
<td>1/16</td>
<td>0.063</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>148</td>
<td>-</td>
<td>-</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>Xgwm272</td>
<td>150</td>
<td>5.5/16</td>
<td>0.344</td>
<td>NE</td>
</tr>
<tr>
<td></td>
<td>152</td>
<td>2/16</td>
<td>0.125</td>
<td>NE</td>
</tr>
<tr>
<td></td>
<td>156</td>
<td>7.5/16</td>
<td>0.469</td>
<td>NE</td>
</tr>
<tr>
<td>208</td>
<td>7/16</td>
<td>0.313</td>
<td>0.232</td>
<td>0.220</td>
</tr>
<tr>
<td>210</td>
<td>2/16</td>
<td>0.125</td>
<td>0.106</td>
<td>0.051</td>
</tr>
<tr>
<td>218</td>
<td>1/16</td>
<td>0.063</td>
<td>0.110</td>
<td>0.108</td>
</tr>
<tr>
<td>226</td>
<td>2/16</td>
<td>0.125</td>
<td>0.035</td>
<td>0.015</td>
</tr>
<tr>
<td>Xgwm312</td>
<td>232</td>
<td>4/16</td>
<td>0.375</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>234</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>236</td>
<td>-</td>
<td>-</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>238</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>240</td>
<td>-</td>
<td>-</td>
<td>0.032</td>
</tr>
<tr>
<td>Xgwm408</td>
<td>150</td>
<td>5.5/16</td>
<td>0.344</td>
<td>NE</td>
</tr>
<tr>
<td></td>
<td>173</td>
<td>-</td>
<td>-</td>
<td>NE</td>
</tr>
<tr>
<td></td>
<td>179</td>
<td>-</td>
<td>-</td>
<td>NE</td>
</tr>
<tr>
<td></td>
<td>181</td>
<td>10.5/16</td>
<td>0.656</td>
<td>NE</td>
</tr>
<tr>
<td>Allele</td>
<td>Parental Status</td>
<td>Allele</td>
<td>Parental Status</td>
<td>Allele Size</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>--------</td>
<td>-----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>183</td>
<td>-</td>
<td>NE</td>
<td>NE</td>
<td>0.002</td>
</tr>
<tr>
<td>208</td>
<td>-</td>
<td>NE</td>
<td>NE</td>
<td>0.004</td>
</tr>
<tr>
<td>210</td>
<td>7/16</td>
<td>0.438</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>222</td>
<td>5/16</td>
<td>0.313</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>224</td>
<td>1/16</td>
<td>0.063</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>226</td>
<td>2/16</td>
<td>0.125</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>228</td>
<td>1/16</td>
<td>0.063</td>
<td>NE</td>
<td>NE</td>
</tr>
<tr>
<td>230</td>
<td>-</td>
<td>NE</td>
<td>NE</td>
<td>0.042</td>
</tr>
<tr>
<td>97</td>
<td>1/16</td>
<td>0.063</td>
<td>0.146</td>
<td>0.121</td>
</tr>
<tr>
<td>109</td>
<td>5/16</td>
<td>0.313</td>
<td>0.271</td>
<td>0.288</td>
</tr>
<tr>
<td>119</td>
<td>1/16</td>
<td>0.063</td>
<td>0.064</td>
<td>0.030</td>
</tr>
<tr>
<td>127</td>
<td>-</td>
<td>-</td>
<td>0.006</td>
<td>0.030</td>
</tr>
<tr>
<td>129</td>
<td>1/16</td>
<td>0.063</td>
<td>0.038</td>
<td>0.061</td>
</tr>
<tr>
<td>133</td>
<td>1/16</td>
<td>0.063</td>
<td>0.146</td>
<td>0.048</td>
</tr>
<tr>
<td>134</td>
<td>4/16</td>
<td>0.250</td>
<td>0.185</td>
<td>0.339</td>
</tr>
<tr>
<td>135</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>137</td>
<td>1/16</td>
<td>0.063</td>
<td>0.096</td>
<td>0.061</td>
</tr>
<tr>
<td>139</td>
<td>2/16</td>
<td>0.125</td>
<td>0.048</td>
<td>0.021</td>
</tr>
<tr>
<td>206</td>
<td>7/16</td>
<td>0.438</td>
<td>0.605</td>
<td>0.512</td>
</tr>
<tr>
<td>208</td>
<td>-</td>
<td>-</td>
<td>0.045</td>
<td>0.08</td>
</tr>
<tr>
<td>210</td>
<td>5/16</td>
<td>0.313</td>
<td>0.207</td>
<td>0.225</td>
</tr>
<tr>
<td>220</td>
<td>4/16</td>
<td>0.250</td>
<td>0.143</td>
<td>0.183</td>
</tr>
<tr>
<td>203</td>
<td>2/16</td>
<td>0.125</td>
<td>0.068</td>
<td>0.156</td>
</tr>
<tr>
<td>205</td>
<td>5/16</td>
<td>0.313</td>
<td>0.266</td>
<td>0.330</td>
</tr>
<tr>
<td>207</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.014</td>
</tr>
<tr>
<td>211</td>
<td>7/16</td>
<td>0.438</td>
<td>0.568</td>
<td>0.365</td>
</tr>
<tr>
<td>217</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.014</td>
</tr>
<tr>
<td>219</td>
<td>1/16</td>
<td>0.063</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td>221</td>
<td>-</td>
<td>-</td>
<td>0.019</td>
<td>0.007</td>
</tr>
<tr>
<td>223</td>
<td>-</td>
<td>-</td>
<td>0.065</td>
<td>0.108</td>
</tr>
<tr>
<td>225</td>
<td>1/16</td>
<td>0.063</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>227</td>
<td>-</td>
<td>-</td>
<td>0.006</td>
<td>-</td>
</tr>
</tbody>
</table>

Status corresponds to parental (P) or novel (N) alleles. Newly arisen alleles are indicated in bold. Distance from parental allele has been estimated from allele size, considering that mutation involved changes in the number of repeats only. NE: non estimated.
Table 2: Mutation rate\(^a\) estimations derived by the four coalescent-related methods (see text) in the population.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Parents-G1</th>
<th>G1-G15</th>
<th>Parents-G15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\mu_1)</td>
<td>(\mu_2)</td>
<td>(\mu_3)</td>
</tr>
<tr>
<td>Xgwm181</td>
<td>2.76</td>
<td>NE</td>
<td>2.64</td>
</tr>
<tr>
<td>Xgwm260</td>
<td>4.97</td>
<td>NE</td>
<td>4.78</td>
</tr>
<tr>
<td>Xgwm312</td>
<td>2.74</td>
<td>NE</td>
<td>2.61</td>
</tr>
<tr>
<td>Xgwm437</td>
<td>1.36</td>
<td>NE</td>
<td>1.29</td>
</tr>
<tr>
<td>Xgwm642</td>
<td>1.36</td>
<td>NE</td>
<td>1.30</td>
</tr>
<tr>
<td>Cfd71_D</td>
<td>4.13</td>
<td>NE</td>
<td>3.94</td>
</tr>
<tr>
<td>Xgwm272</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Xgwm408</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Xgwm427</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

\(^a\)10\(^{-3}\) ; NE : non estimated, NA : non available
Figure legends

Figure 1: Experimental design for the population studied. The four generations of crosses were performed manually; bulk multiplication and evolution under natural selection were conducted under the natural mating system of wheat (mainly selfing); the population was grown one generation per year with a gap of three years before sowing G0 seeds; one generation starts in the autumn of year $n$ and is completed in the summer of year $n+1$.

Figure 2: Gene diversity in the population (16 parents and G15), and in the wheat collection from the INRA Clermont-Ferrand.
Anne-Laure RAQUIN, Frantz DEPAULIS, Amaury LAMBERT, Nathalie GALIC, Philippe BRABANT, Isabelle GOLDRINGER Figure 1
Anne-Laure RAQUIN, Frantz DEPAULIS, Amaury LAMBERT, Nathalie GALIC, Philippe BRABANT, Isabelle GOLDRINGER Figure 2