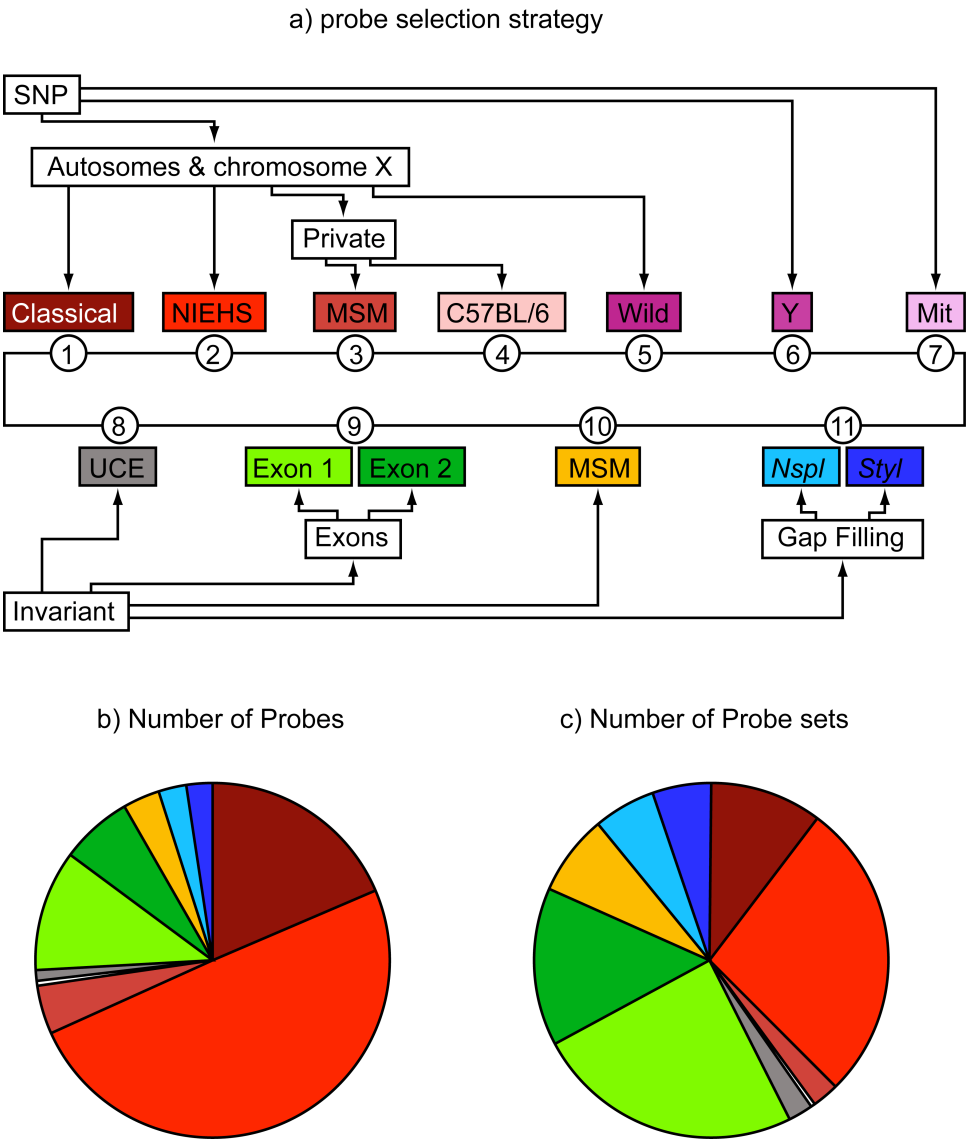
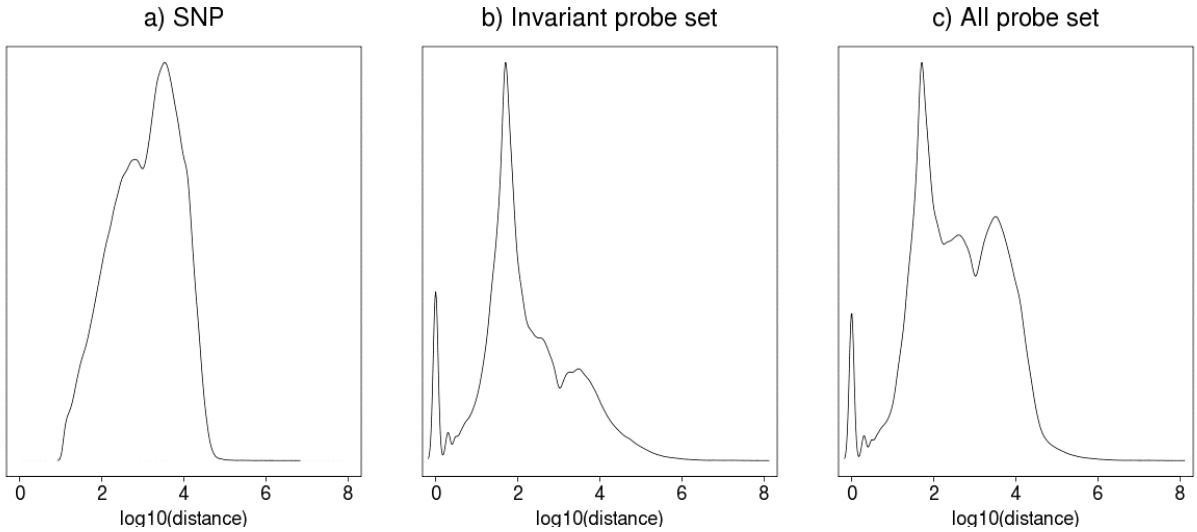
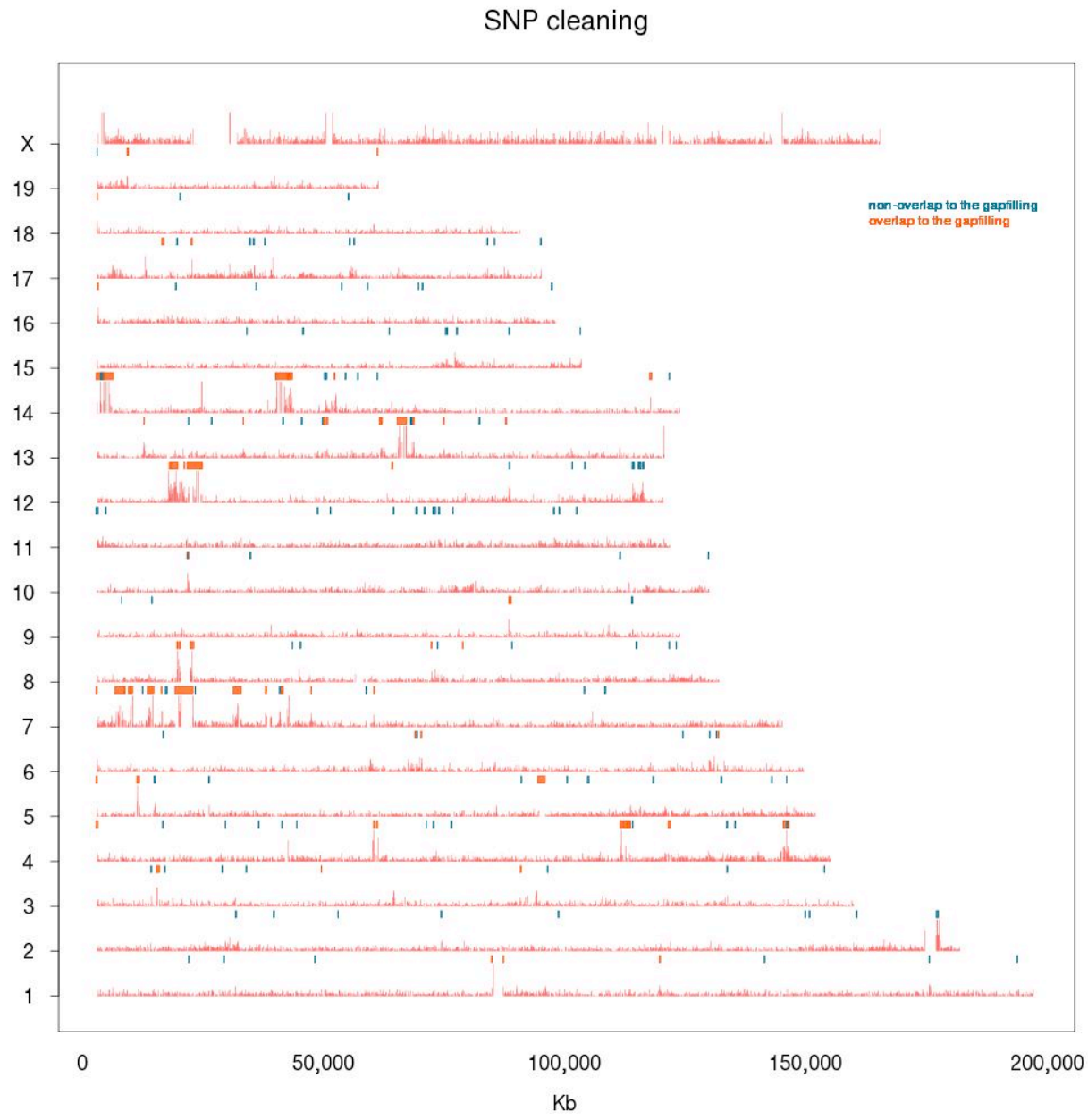
Supplementary Material



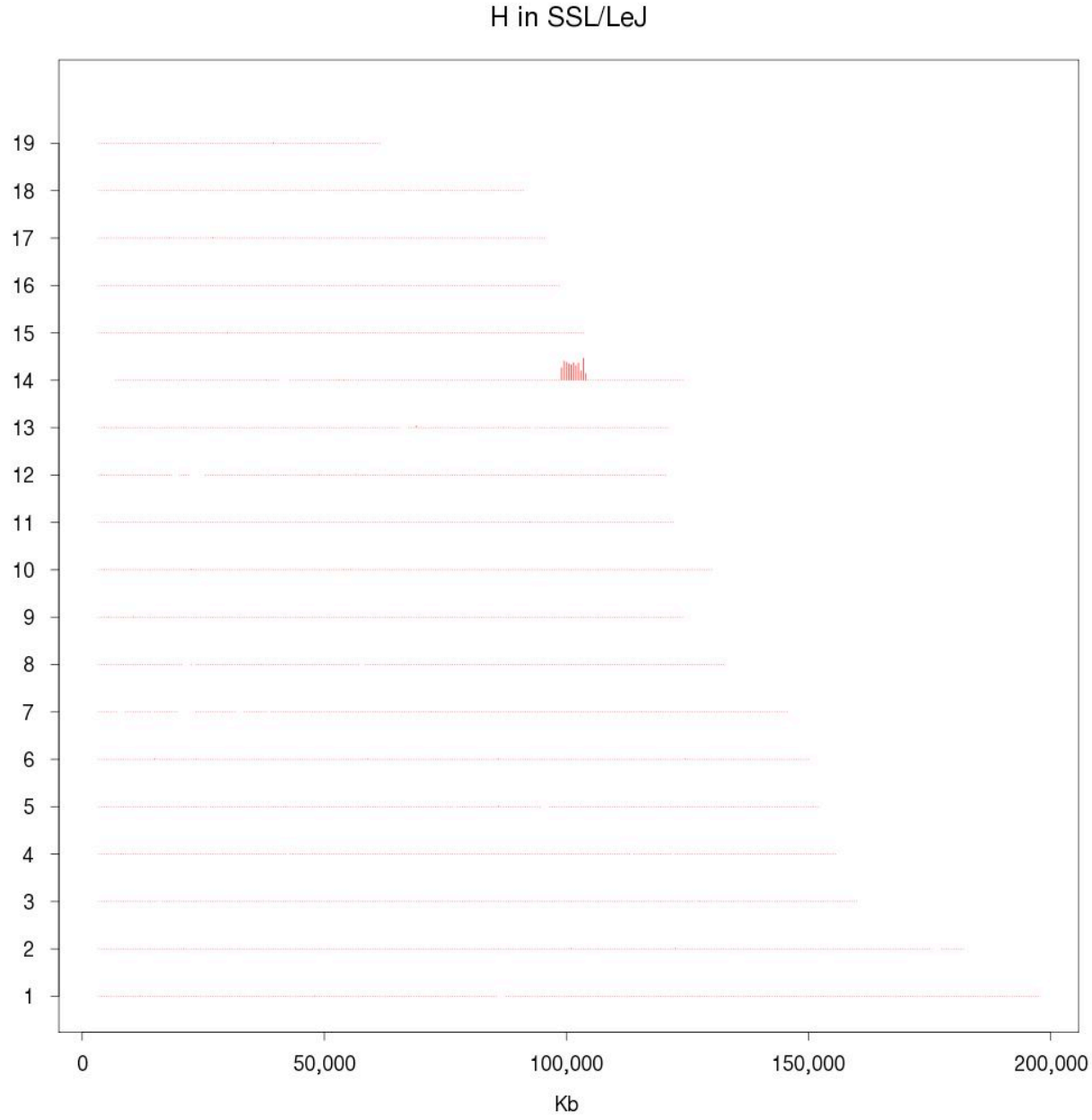
Supplementary Figure 1 : **Design strategy and content of the Mouse Diversity array**. *a)* Probe selection strategy is explained in detail in Box 1 (SNPs) and Box 2 (Invariant). *b)* Number of probes in the final array selected on the basis of each of the strategies shown above. *c)* Number of probe sets in the array.



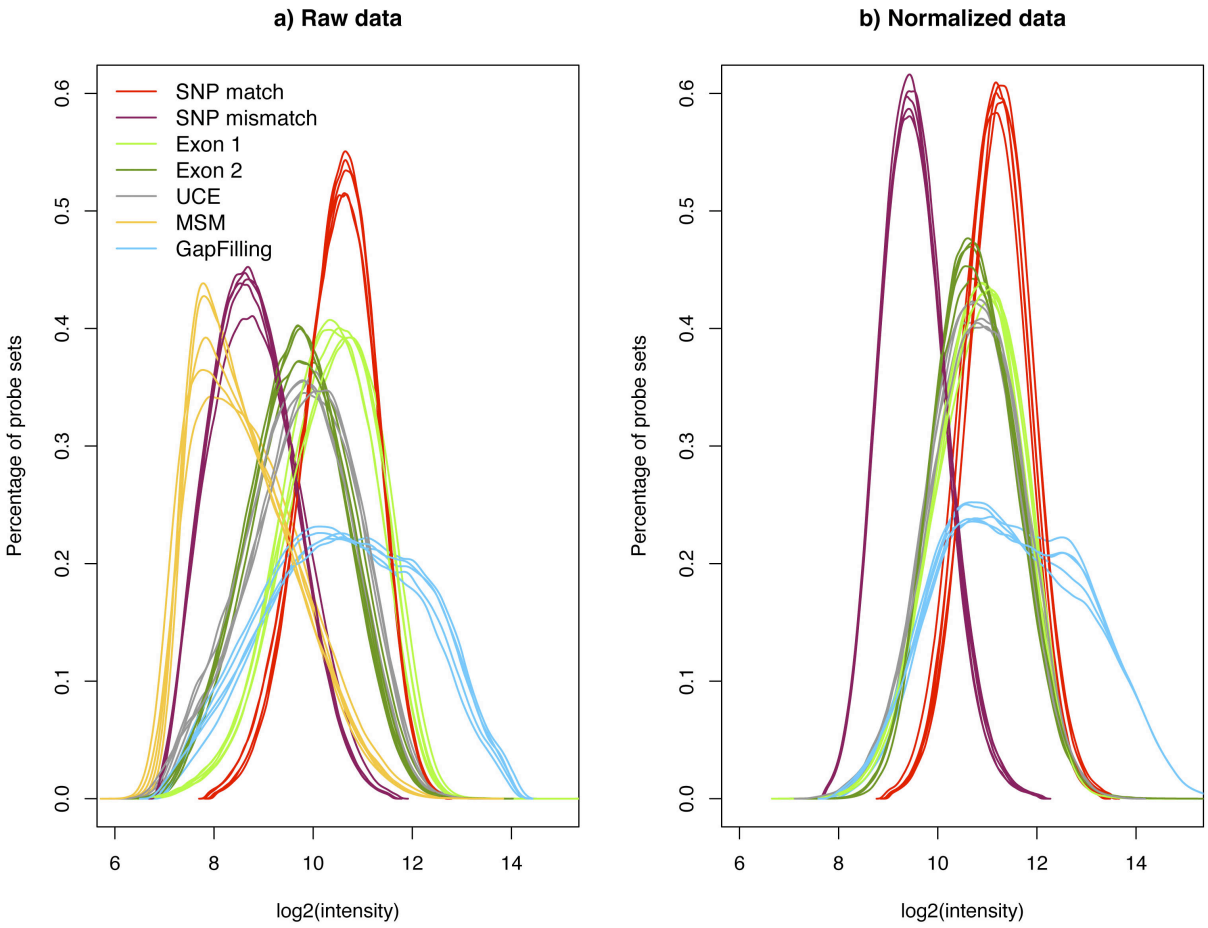
Supplementary Figure 2 : Distribution of the distance between consecutive probe sets.



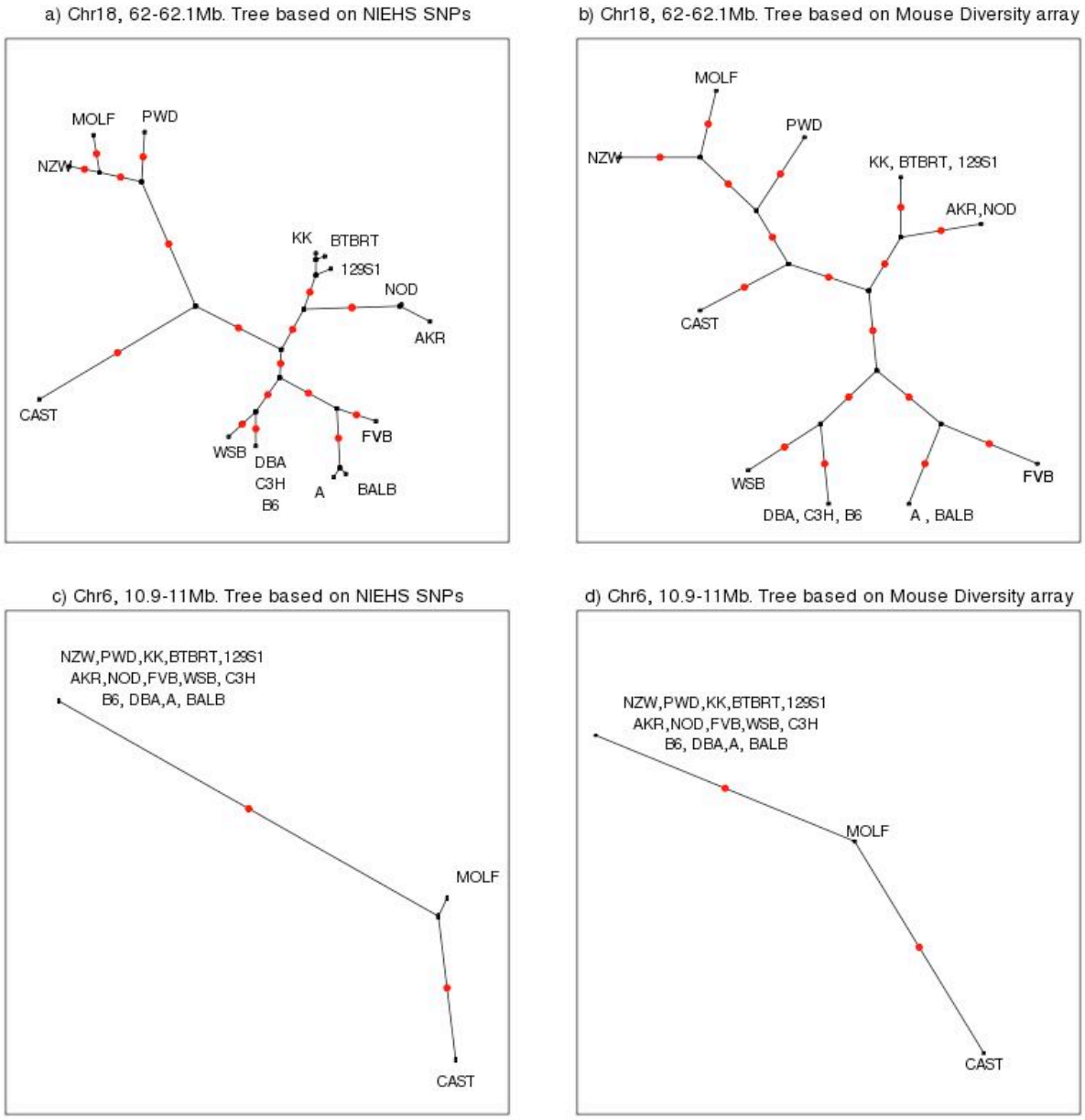
Supplementary Figure 3 : Poorly performing SNPs and SNPs in potential segmental duplicated region.



Supplementary Figure 4 : Residual heterozygosity in the SSL/LeJ strain in 200kb windows are indicated as vertical red bars.

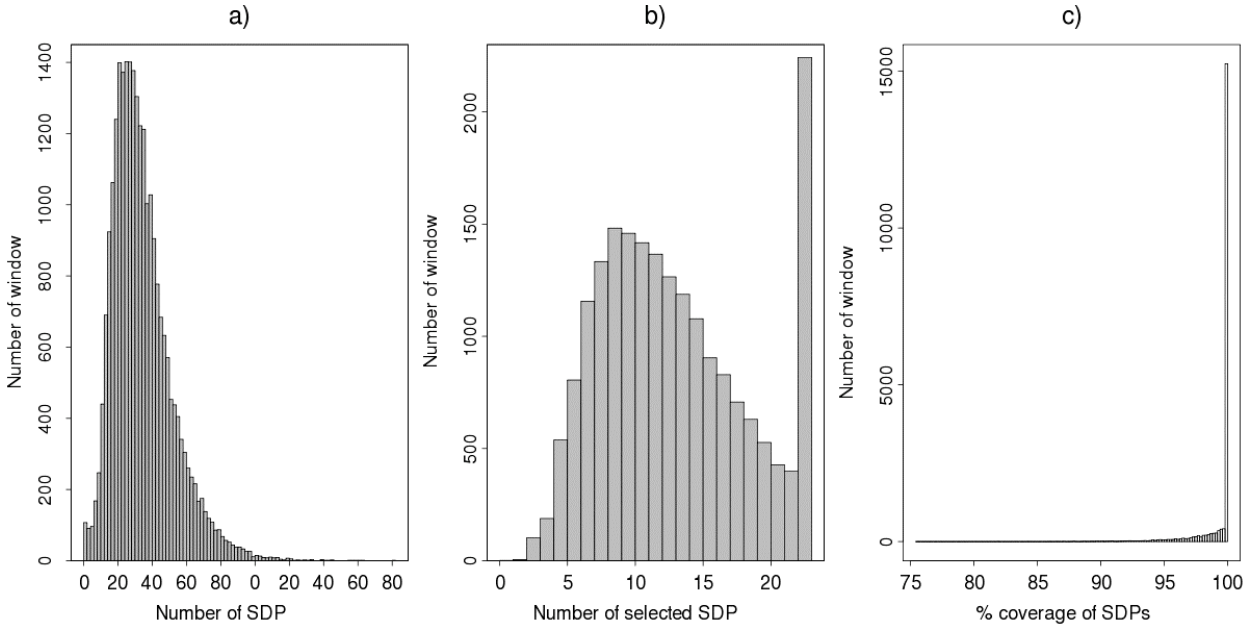


Supplementary Figure 5 . Intensity distributions for different types of probes based on hybridization to C57BL/6J. We divided the probes into seven classes, two based on whether the allelic variant present in SNP probes is a match or mismatch for the C57BL/6J sequence and five based on the main strategies for invariant probes. Density curves are standardized to have total area one and thus are not proportional to the numbers of probes in each class. Density curves are show for five classical strains, C57BL/6J, BALB/cByJ, AKR/J, A/J, FVB/NJ. a) Raw intensity data. b) Normalized intensity data. Normalization was obtained by fitting $y = x + f_1$(length of NspI fragment) + f_2(length of StyI fragment) + f_3(G+C content of probe), where $f_i()$ are cubic splines with 3 df, y is the observed intensity and x is the normalized intensity. MSM probes were excluded from normalization because the fragment lengths are not known.



Supplementary Figure 6 . Selection strategy for SNPs from the NIEHS data. Phylogenetic tree from chr18, 62M-62.1M (*a and b*) and chr6, 10.9M-11M (*c and d*). *a and c)* Phylogenetic trees based on the complete SNP data for each interval. Each branch represents a strain distribution pattern (SDP) and the branch length is proportional to the frequency of the SDP in that interval. *b and d)* Phylogenetic trees based on the NIEHS SNPs present in the array. SNPs representing the most frequent SDPs were selected until 98% of the local variation was covered. SDPs selected for

inclusion in the array are show with a red dot. In tree a 17 SDPs were selected while in tree c only two SDPs were selected. One SNP per SDP was selected, thus in the trees based on SNP present in the array, all the branches from have the same length.



Supplementary Figure 7 : Strain distribution patterns (SDPs) in each 100 kb interval (a), number of selected SDP per interval (b), and genome coverage using up to 22 SDP (c). Among 25,393 intervals that we considered, 20047 intervals had more than 20 SNPs with more than 30% of complete SNPs. If we select up to 22 SDPs in each interval, they covers more than 98% of genome in 90% of intervals and on average 95% of genome for the remaining 10% (or 2242) of intervals.

Selection Categories	Window Size	Selection Criteria	Iterations	Type	SNP Number
Classical Inbred Strain	40 kb	High MAF	3	SNP	187,256
NIEHS	100 kb	One SNP/SDP until 98% of SNPs are represented	Shift 50 kb	SNP	483,431
C57BL/6J Singletons	1 Mb	A/J, DBA/2J, 129S1, MSM/Ms	3	SNP	3,376
MSM/Ms Singletons	100 kb	C57BL/6J, A/J, DBA/2J, 129S1	3	SNP	49,930
Wild	na	*M. macedonicus, M. spretus, M. spicilegus* and *M. musculus* subspecies	na	SNP	990
Chr Y	na	All	na	SNP	83
Chr M	na	All	na	SNP	20
Total					725,086

Supplementary Table 1 : SNP selection criteria in the test array. We set seven selection criteria fit for our purpose, and Table shows the selection criteria, and number of SNPs in each categories. The selection process yielded 725,086 SNPs that were randomly divided and printed on two test arrays. The test arrays were used to eliminate poorly performing SNPs and to select the best performing probe for each SNP included in the final array. For each SNP, 16 to 18 probes were printed on the test array including perfect matches for each allele in both the forward and reverse strand. Probe sequences were offset 0, 2 or 4 bp from the central SNP position to allow for the selection of the best performing probe. The two test arrays were hybridized with 82 and 87 DNA samples, respectively. We used the BRLMM-P algorithm to obtain genotype calls (see **Methods**). We scored each SNP based on the number of complete genotype calls and concordance rate compared to genotypes in public databases. In the test arrays the average call rate for all SNPs was 93.63%, and concordance rate with SNPs reported in public resources was 88%. We retained 623,124 SNPs for the final array, as summarized in **Table 1** and **Figure 1**. For each selected SNP, we identified one best performing probe on each strand. We used the silhouette score[19] to assess genotype class separation and we used variance of intensity as a measure of within cluster consistency to guide probe selection. The plus and minus strand probes may have different offsets relative to the

targeted SNP. Probes for each allele in each strand at each SNP were printed in duplicate on the final array, resulting in a total of 8 probes per SNP assay.

Strain	Call rate	Het rate	Known Genotype	concordance Rate
129SvlmJ	99.642	1.319	0.94	0.995
A/J	99.602	1.1	0.972	0.998
AKR/J	99.712	1.177	0.928	0.997
BALB/cByJ	99.718	1.072	0.927	0.998
BTBRT+tf/J	99.536	1.308	0.926	0.995
C3H/HeJ	99.365	1.201	0.927	0.996
C57BL/6J	99.463	0.897	0.999	0.999
DBA/2J	99.59	1.218	0.965	0.997
FVB/NJ	99.714	1.274	0.927	0.996
KK/HIJ	99.456	1.626	0.926	0.994
NOD/LtJ	99.463	1.337	0.926	0.996
NZW/LacJ	99.486	1.466	0.928	0.996
WSB/EiJ	98.041	2.445	0.924	0.992
PWD/PhJ	98.155	5.262	0.909	0.992
MOLF/EiJ	97.688	5.451	0.912	0.991
MSM/Ms	96.966	6.032	0.206	0.985
CAST/EiJ	96.306	6.871	0.909	0.982
SPRET/EiJ	92.386	14.797	0.14	0.986
PANCEVO/EiJ	92.092	12.422	0	0

Supplementary Table 2 : Performance of SNP probes.