SUPPLEMENTARY MATERIAL

**Table S1**. List of samples used in this update of MiniMUGA. The table provides a randomly assigned ID, the laboratory source of the samples, the type of sample regarding the level of inbreeding, a reference for previously published samples, whether the sample was used in reclustering exercise, used in building the consensus genotypes for a given strain or serve as positive or negative control for the two new constructs added to the report. This is described in more detail in **Table S3**.

**Table S2**. Marker annotation updates. This file includes several annotations for each marker including chromosome, position (in GRCm38), name, probe, rsid, tier\_2022, diagnostic, partial\_diagnostic, diagnostic\_allele, construct\_info, positive\_threshold, negative\_threshold, and recluster. These annotations are described in more detail in **Table S3.**

**Table S3**. This table provides a detailed description for each column or field listed in **Table S1** and **Table** **S2.**

**Table S4:** This table includes consensus genotype calls for 242 inbred strains and substrains at 10,819 SNP markers. For easy reference, it also duplicates some marker annotations from **Table S2** related to marker performance, diagnostic capability, and homoplasy. The strain name column headers also include the number of biological replicates (in parenthesis) used in the consensus for that strain. This parenthetical can and should be safely parsed out for most usages.

**Table S5**. The rows in this table list all classical inbred strains with consensus genotypes, while the columns list the strain groups used for determination of diagnosticity of SNPs plus a common outgroup that includes all classical inbred strains without diagnostic SNPs. Only cells with numbers or N/A should be considered. Numbers in each cell are the number of diagnostic SNPs followed by the number of partially diagnostic SNPs for each substrain.

**Table S6.** Selected examples of the effect of the new genome analysis pipeline in improved estimation of the contribution of the primary and secondary backgrounds. The table lists the percentage contribution of primary, secondary, heterozygous, and unknown backgrounds in 23 samples under the original and updated algorithms. Differential is the difference between new and old estimates for each contribution.

**Figure S1.** Examples of clustering and consistency of calls for SNPs annotated as Tier 1A, 1B, 1C, 2A, 2B, 2C and 4 are shown. In Illumina Infinium genotyping platform the X axis shows the normalized fluorescence intensity for A and T alleles while the Y axis shows the normalized fluorescence intensity for C and G alleles. Homozygous genotypes are shown as open circles, heterozygous as blue circles, and N (no calls) as red filled circles. The Tier 1A marker gJAX00183552 (rs29098241) there are three tight and well separated clusters, and 8,565 samples have the expected A, C, or heterozygous genotypes; two samples with expected homozygous T genotypes did not return a genotype (N). For Tier 1B marker SBJ190768497 (rs248647047) with observe the expected three clusters but there are 51 samples with no calls (N). For Tier 1C marker gUNC1978827 (rs246827811) the discrimination between the clusters is poor and there are samples with incorrect genotype calls (for example samples incorrectly called homozygous T in the heterozygous cluster). Tier 2A shows marker gUNC29908079 (rs37961860) with the three expected clusters (ref, alt and het) and an additional cluster near the origin, all ref, alt, and het genotypes are consistent with the clustering. In Tier 2B marker gUNC16481123 (rs30469782), there are samples that likely have the A, G or het genotypes that classified as N, but no sample is genotyped incorrectly. Note also that most of samples in the cluster at the origin likely have the “G” genotype plus one or more additional off target variants within the 50 bases long probe in present in the array. Tier 2C marker gUNCHS025276 (rs37051043) has similar issues to the previous panel but there are several samples that are genotyped as “G” but are in all likelihood heterozygous for the G allele and the off-target allele close to the origin. Finally, the figure shows two examples of SNP classified as Tier 4, gUNC569059 attempts to cluster the samples but the resulting clusters are unresolved and not useful, while gJAX00225545 is monomorphic and thus could eventually be rescued if samples the T genotype exist.

**Figure S2.** Intensity of individual probes for the two new constructs detected by MiniMUGA, Chicken HS4 insulator (cHS4) and Flippase (Flp). For each construct, individual samples are plotted as dots with vertical positions reflecting the intensity values for seven construct markers in that sample. The samples are classified as negative controls (left), experimental (center), and positive controls (right). The dot color denotes whether the sample is part of the negative control set (blue), the positive control set (red), or an experimental (gray) for the respective construct. Note that for Flp probes that are two samples among the positive controls with consistently higher intensity for all individual probes. This is likely due to an increase in the number of copies for this construct in these samples. For cHS4, there are two experimental samples with also higher intensities for four probes and zero intensity for the other three probes (ZZcHS40020, ZZcHS40024, and ZZcHS40026). This is likely due to a combination of off-target variation in the negative probes and an increased copy number of the construct itself in some samples. This type of information could be used in the future to discriminate between versions of the cHS4 construct. Note that these two experimental samples are considered positive overall (Fig 2).

**Figure S3.** SNPs with diagnostic alleles and homoplasy. The two panels show the same global phylogenetic tree and nine SNPs with different segregation patterns across 14 mouse strains (A to H) and two sets of three substrains (A1 to A3 and D1 to D3). The strains are classified as classical (A to F) or wild derived (G to H). Panel A shows the timescale in natural years while panel B shows the timescale in log10. Nine SNPs are shown as circles with alternative alleles as open or color filled circles. The figure shows the branches where diagnostic SNPs arise in red for group A and in blue for group D. The green vertical dashed line represents the start of inbreeding of classical inbred strains. The vertical blue dashed line denotes the initial time of independent breeding and genetic drift between D substrains. The vertical red dashed line denotes the initial time of independent breeding and genetic drift between A substrains. The figures show whether each one of nine SNPs has diagnostic alleles according to the definition in the text, is homoplasic with the second mutation present in one or more wild-derived strains, is homoplasic with the second mutation present in one or more classical strains. Finally, the figures show whether a SNP is annotated as diagnostic in **Table S2**. SNPs 6 and 9 are diagnostic but are not annotated as such because the second event is present in a classical inbred and thus may lead to false positive identification of a substrain. SNP 5 is shown as a question mark for homoplasic wild-derived because the uncertainties created the complex population structure and subspecific introgression of classical inbred strains (Yang et al. 2011).

**Figure S4**. The figure shows an example of the current layout of the MiniMUGA report for an incipient congenic sample from the MMRRC catalog. This is one of two MMRRC samples from the B6.129P2-*Gabraqtm1Dgen*/Mmnc (RRID:MMRRC\_011620-UNC) strain genotyped with MiniMUGA. The primary background is a mix of at least two C57BL/6 substrains and the secondary background is one of several 129 related strains including 129P2/OlaHsd, the strain listed as donor by the strain developer. Note that the inbreeding estimate (99.5%) is based in the segregating interval for primary and secondary backgrounds (13 Mb on chromosome 19) and does not include segregation for the two C57BL/6 substrains, C57BL/6J and C57BL/6NCrl. Based on the fraction of heterozygous calls listed in the Backgrounds Detected section, the level of heterozygosity at the substrain level is considerable.

**Figure S5.** Example ideograms of F1 hybrids. Panel A shows an example ideogram for a F1 between two related substrains, a (BALB/cByJxBALB/cJ)F1 female (ID, SB1956). Panel B shows the ideogram for a (DBA/1LacJxC3H/HeNTac)F1 male (ID, JG5569). Note that the mitochondria and Y Chromosome are returned as IBD. Finally, there are open triangles for all for substrains, 12 for BALB/cByJ, seven for BALB/cJ, eight for DBA/1LacJ and 12 for C3H/HeNTac, reflecting the presence of partially diagnostic SNPs.

**Figure S6.** Two examples of fractured ideograms. Panel A is a sample with a mix of more than three backgrounds, C57BL/6J, C57BL/6NCrl, 129S2SvHsd and at least of more background present on chromosomes 13 and 14. Panel B is an outbred sample on an outbred background. Both are MMRRC samples.