TWIST
BIOSCIENCE

Twist 96-Plex High Throughput NGS Library Prep for Genotyping in Populations

Azeem Siddique,[1,3] Gaia Suckow,[1,3] Nils Homer,[2] Alvaro Gonzalo Hernandez,[4] Phillip Ordoukhanian,[1,3] Steve Head,[1,3] Keith Brown[3], Lior Glick,[5] Kobi Baruch,[5] Paul Doran[5]

[1]The Scripps Research Institute, La Jolla, CA; [2]Fulcrum Genomics, Somerville, MA; [3]Twist Bioscience, South San Francisco, CA; [4]University of Illinois at Urbana-Champaign, Urbana, IL; [5]NRGene, Israel

ABSTRACT

High throughput genotyping technologies are required for large-scale population genetics. Evolutionary biology studies, human disease research and large-scale agricultural breeding programs all lend themselves to technologies that are able to provide more information at lower cost. Over the past decade, genotyping technology has transitioned from PCR-based SNP assays to microarrays, and is now shifting toward high-throughput genotyping by sequencing (GBS). The Twist 96-Plex Library Preparation Kit allows for the preparation of NGS libraries from up to 960 individually barcoded samples in a few hours with automation. When combined with low coverage sequencing and imputation-based genotype analysis, the result is an order of magnitude greater information at a significantly reduced cost. Here we present data on 96 Zea mays (maize) samples consisting of 4 parent populations and 92 recombinant inbred lines (RILs). For each sample, hundreds of thousands to millions of haplotype markers, including SNVs and structural variants, are accurately detected. A minimum of 95% complete coverage of direct and imputed markers is obtained for each RIL. The approach can be applied to any species, regardless of genome size or GC content. In this study, a median of >1 million markers were genotyped by sequencing on an Illumina HiSeq 4000 instrument for an estimated cost of library construction and sequencing of < $25 per sample.

OVERVIEW

Samples were obtained from a public cornfield in Iowa. Parental lines include B73, PH207, LH82 and PHG39. RILs include B73 x PH207 (n=16), B73 x PHG39 (n=40), B73 x LH82 (n=16) and LH82 x PH207 (n=20). Plant material was sent to LGC US laboratory (Beverly, MA) for DNA extraction. DNA concentrations ranged from 7.6 – 18 ng/mL with a mean of 12 ng/L. All samples were treated as though the DNA concentrations were 12 ng/mL. 100 ng of DNA was used as input into the RipTide library prep" with "Twist 96-Plex library prep (see Figure 1). In a separate, virtually identical experiment, 4 µL of each sample were used as input in the same library prep protocol. Libraries were prepared and sequenced at the University of Illinois Roy J. Carver Biotechnology Center (Urbana, IL). The 100 ng input samples were sequenced on 3 lanes of a HiSeq 4000 flow cell using 2 x 150 bp paired-end chemistry. An additional lane on the flow cell was used for the 4 µL samples. Data was de-multiplexed (Fulcrum Genomics) and uploaded to the GenoMAGIC™ pipeline (NRGene) for haplotype calling and genotype imputation.

SCHEMATIC OF THE TWIST 96-PLEX WORKFLOW

A) Random primers with 5' barcoded Illumina adapter sequences are annealed to a denatured DNA template (Figure 1). A polymerase extends each primer, generating a copy of the DNA template. Polymerization is terminated with a biotinylated dideoxynucleotide of which there is a small fraction in the nucleotide mix. **B)** Primer-extended products are captured on streptavidin-coated magnetic beads. The beads are washed to remove excess reactants. **C)** A second 5' adapter-tailed random primer is used with a strand-displacing polymerase to convert the captured DNA strands to a dual adapter library. **D)** The beads are washed once again to remove excess reactants and displaced primer-extended products. **E)** PCR is used to amplify the products and add an index barcode. In the high throughput version of the library prep, individual samples are uniquely labeled in a 96-well plate with the use of a uniquely barcoded random primer in each well of the plate. After the initial labeling step, products from all wells are pooled and all subsequent steps are performed in a single tube. An optional plate barcode is added during the PCR step to allow for multiple 96-sample plates to be sequenced simultaneously.
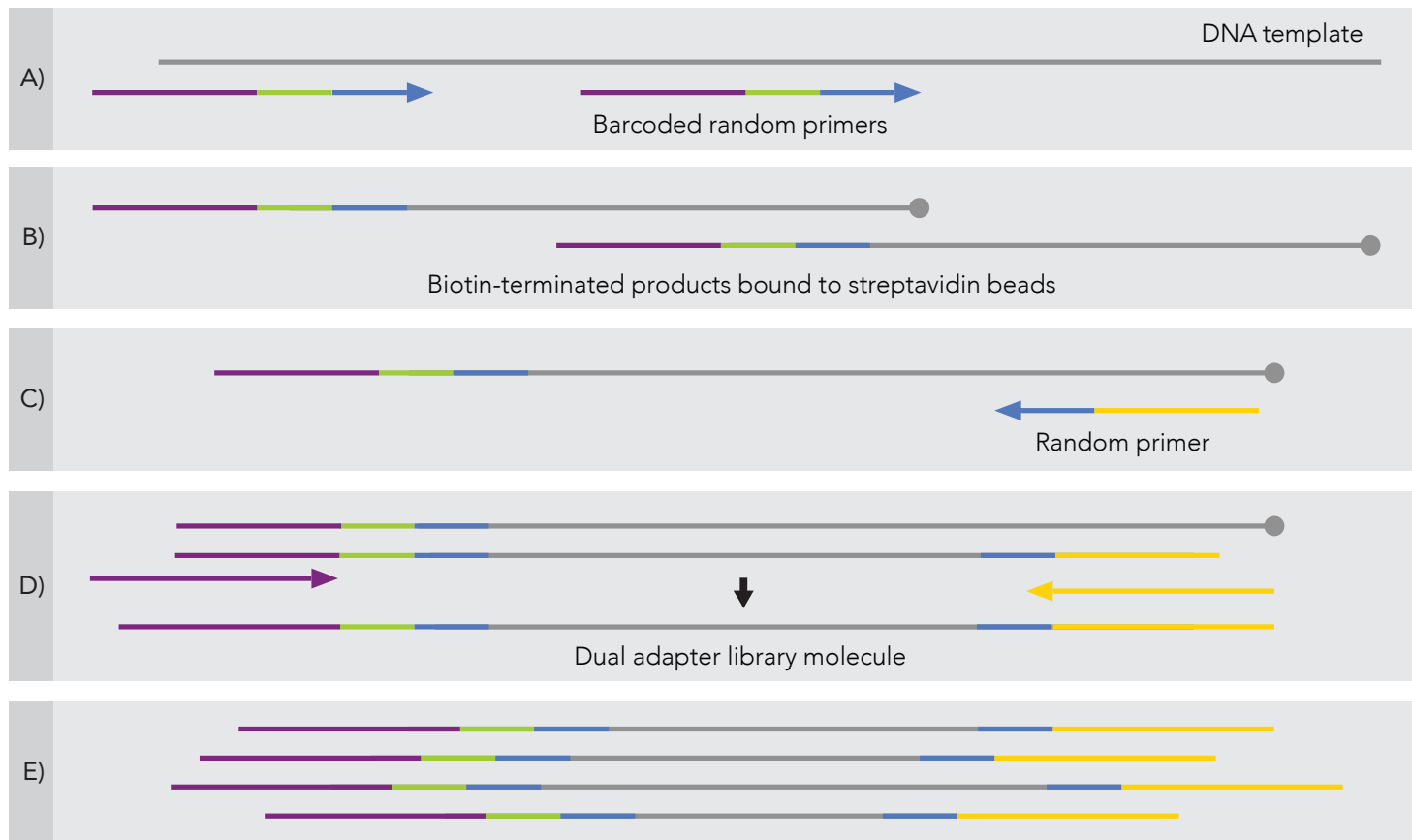

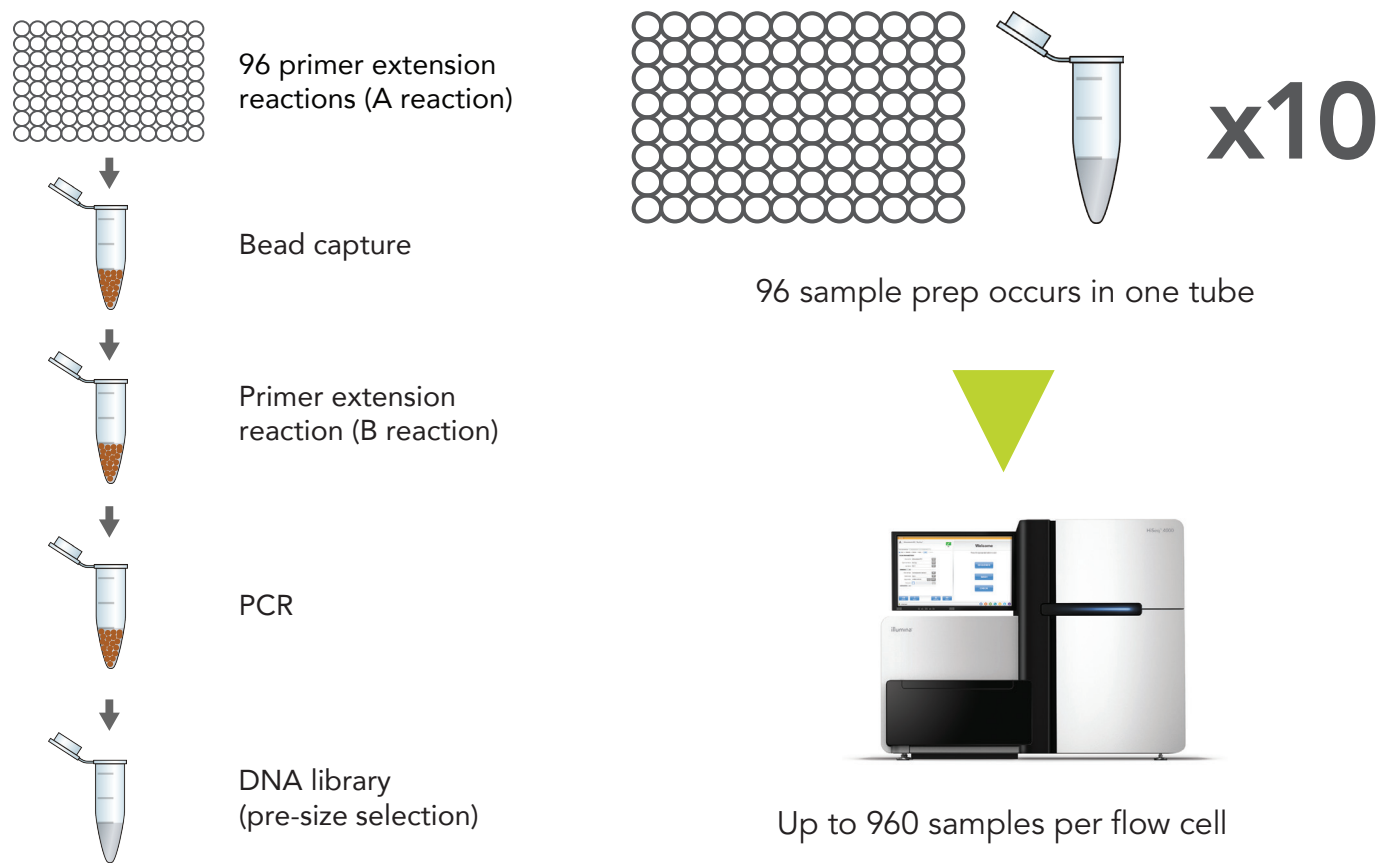**Figure 1:** Schematic of the Twist 96-Plex workflow

Figure 2: High throughput workflow

RESULTS

Raw sequence data was of high quality with mean Q scores > 30 for both Read 1 and Read 2 (Figure 3).

Mapping rates across all samples ranged from 92%–97% with a median of 95%. Median read counts were 22.4 million and 7.1 million for the 100 ng and 4 µL samples (3 lanes versus 1 lane of the flow cell) for an estimated genome coverage of 1.4X and 0.44X, respectively (assuming a 2.4 Gb genome size).

Results from the NRGene GenoMAGIC haplotype analysis show that hundreds of thousands to millions of haplotype markers were detected per sample. A median of >1.057 million haplotype markers were detected for the 4 µL samples and a median of >1.9 million haplotype markers were detected for the 100 ng samples. Haplotype markers are evenly distributed across genomes. Figure 4 below shows the distribution of markers across chromosome 1 of the B73 parental sample (100 ng input). To compare this genotyping method to a known genotyping method in maize, we compared the haplotype markers detected by GBS to the Affymetrix Axiom 616K maize genotyping SNP marker set. The number of Axiom SNP markers detected by GBS in each RIL population is presented in Table 1.

Based on haplotype marker profiles, the GenoMAGIC platform can infer similarity between samples. Figure 5 shows a graphical display of the similarity inference for chromosome 1 of the B73 x PH207 RIL population.
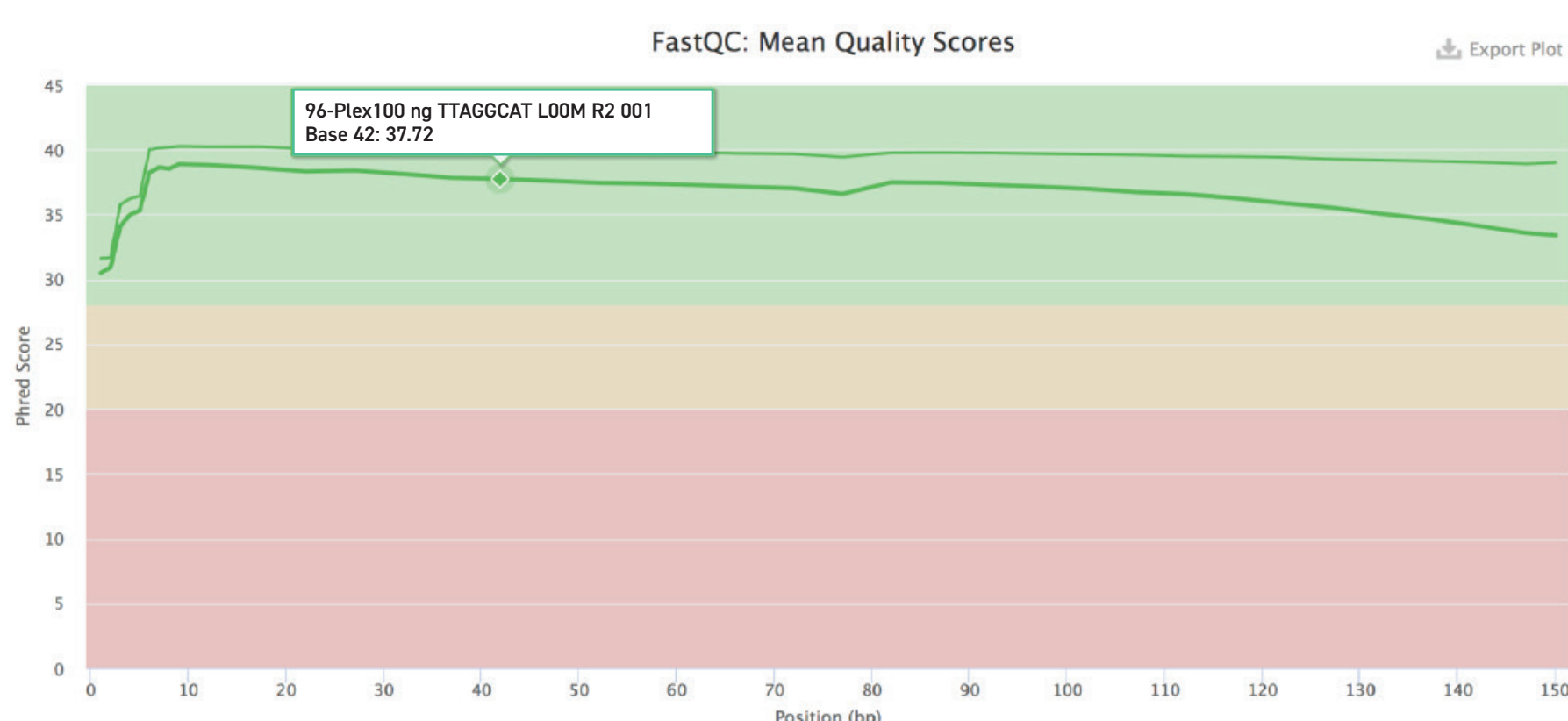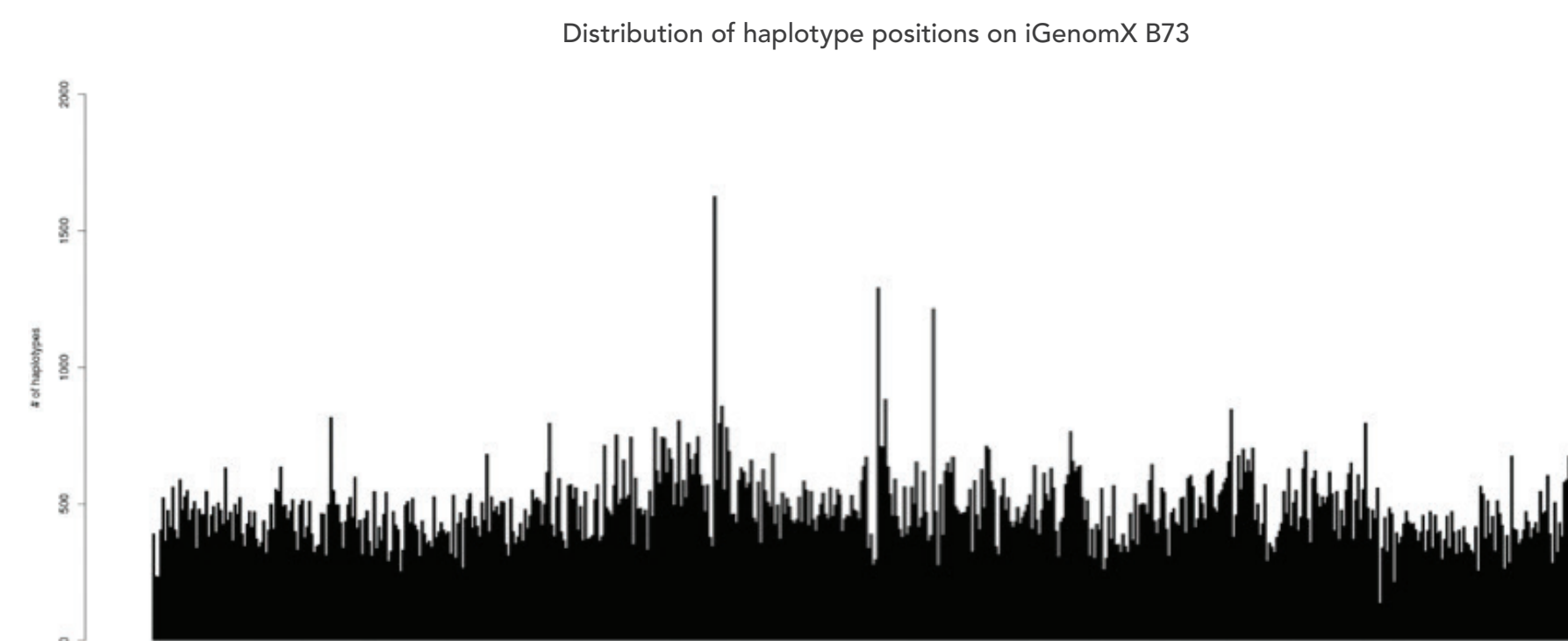

Figure 3: Twist 96-Plex mean quality scores

Figure 4: Distribution of markers across chromosome 1 of the B73 parental sample (100 ng)
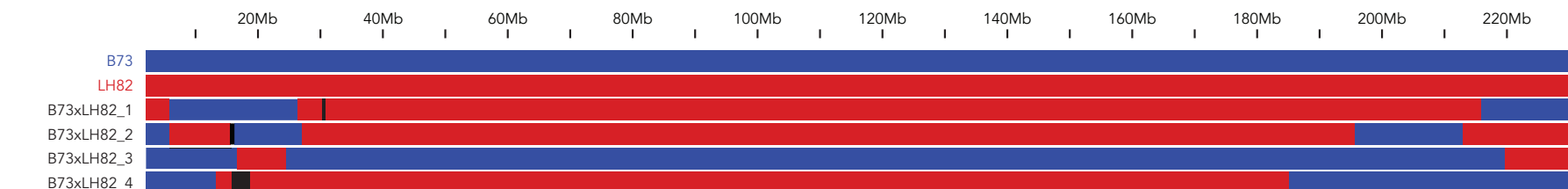

Figure 5: Haplotypes similarity across chromosome 2 of four RIL lines belonging to a B73 x LH82 segregating population. Blue and red bars represent B73 and LH82 haplotypes, respectively. Black bars represent regions of uncertainty where similarity could not be reliably inferred.
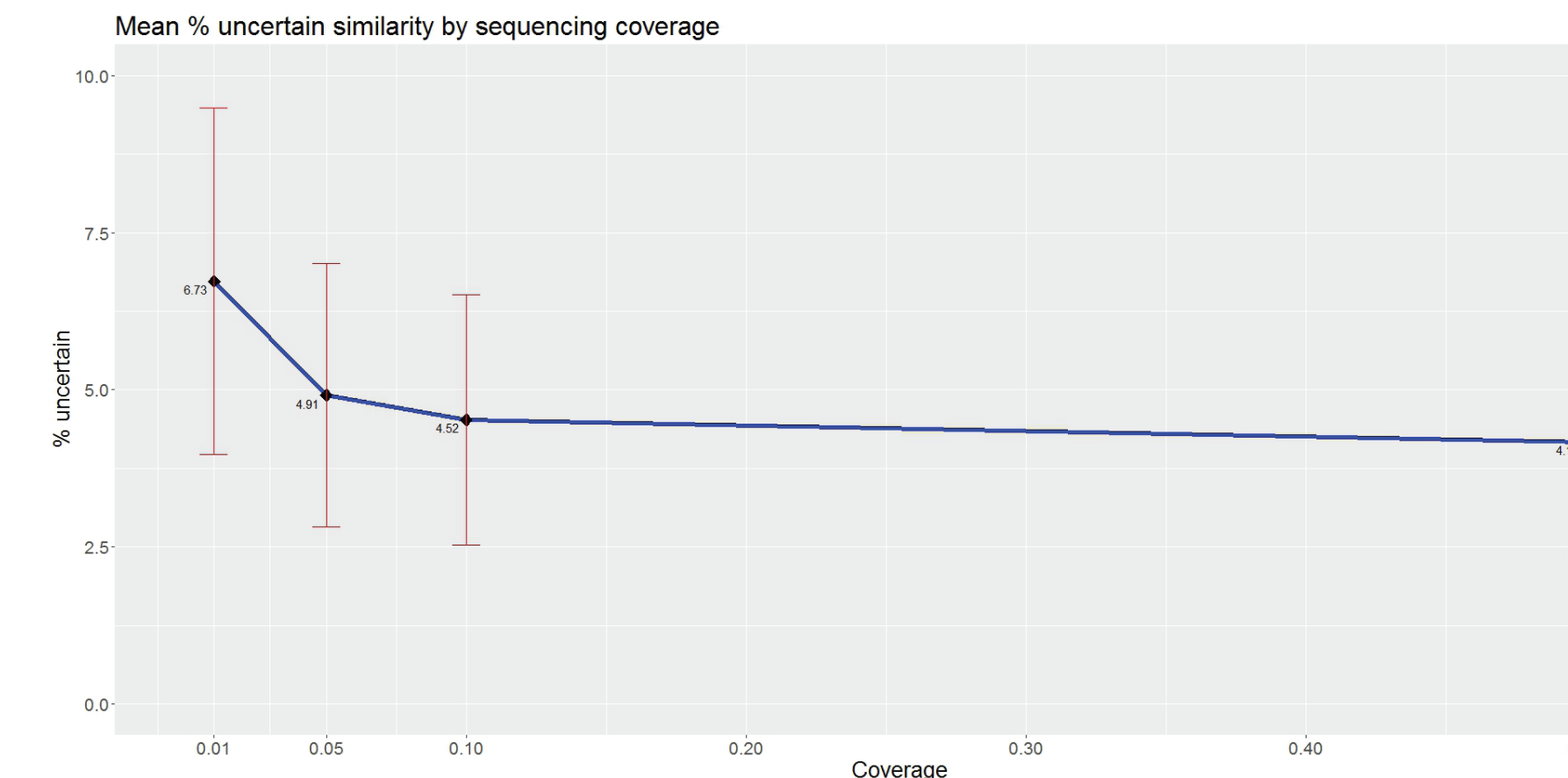

Figure 6: Relationship between sequencing coverage and the % of genome for which similarity could not be reliably inferred. Points represent the coverage depths used in the experiment. Data was collected from the whole B73 x LH82 RIL population (16 samples-mean and SE shown) and summarized across the whole genome.

POPULATION	P1 # OF MARKERS	P2 # OF MARKERS	NUMBER OF PROGENY	AVERAGE MARKERS CALLED PER PROGENY	AVERAGE ERROR RATE (%)
b73xlh82	607,672	603,109	16	576,955 ± 13,058	0.84%
lh82xph207	603,109	608,716	20	583,543 ± 6,854	0.82%
b73Xphg39	607,672	554,081	40	563,232 ± 3,410	0.82%

Table 1: Average number of SNPs markers (out of the 609,442 possible on the Axiom 616K) detected in each of the RIL populations.

CONCLUSIONS

- Twist 96-Plex library-generated data is of high quality and reads are distributed uniformly across the genome. Uncovered regions are small and the overall ability to infer haoplotypes is high.

- The GenoMAGIC pipeline enables the sequence imputation of haplotype in RIL progeny with >99% accuracy and to >94% completeness in every sample (based on 0.01X coverage).

- Translating the haplotypes to SNP markers allow more than 92% of the Axiom 616K SNP set to be imputed successfully.

- The combination of the cost-effective Twist 96-Plex library prep, ultra low-coverage sequencing and high-resolution haplotype imputation offered by NRGene enables high quality, high throughput genotyping at low cost.