SUPPLEMENTARY INFORMATION

SUPPLEMENTARY METHODS

COMPARATIVE GENOMIC HYBRIDIZATION (CGH)

The construction of a microarray for comparative genomic hybridization (CGH) based on the assembled ESTs for Copy Number Variation (CNV) analysis has been described previously (Vishwanathan et al., 2017). Since the original manufacturer of the array no long provide the oligoDNA array, a new 4-plex array each with 167,508 DNA probes was custom designed and manufactured by Agilent. The array focused on the transcript coding regions of the genome.

For CGH array hybridization, genomic DNA was extracted from cells using DNeasy Blood & Tissue Kit (Qiagen, Vaencia, CA) or NucleoSpin® Tissue Kit (Macherey-Nagel, Duren, Germany). DNA was quantified using a NanoDrop 1000 Spectrophotometer and the genomic DNA integrity was assessed in a 1% agarose gel. The DNA pre-processing, labeling and hybridization to the Agilent microarray was done by the genomics core facility at the University of Iowa. Briefly, the DNAs were labeled using the SureTag DNA Labeling Kit followed by hybridization to a 4x180K 2-dye expression microarray from Agilent (Design ID 075155) as per manufacturer's instructions. The arrays were scanned using Agilent SureScan Microarray Scanner G2600D. The raw intensity data from the microarray hybridization images were obtained from Agilent CytoGenomics 3.0 software in form of a .txt file and was normalized by LOWESS method. Subsequent analysis was performed using MATLAB® (version 2015b; MathWorks). Further, a circular binary segmentation-based DNAcopy algorithm was used to identify segments affected by copy number variation (Olshen et al., 2004; Venkatraman & Olshen, 2007).

INTEGRATION SITE ANALYSIS

Sequence Capture Method

Integration site analysis for rDG_IgG was performed using sequence capture. A diagram of this process is shown in Supplementary Figure S3. Genomic DNA (gDNA) was sheared, and a size selection was performed to isolate DNA in the 300-500bp range. An indexed and universal adapter were ligated to either end of the sheared fragments. The Agilent SureSelect Enrichment System was then used to enrich for fragments containing vector sequence. Short biotinylated RNA probes (120bp) were designed to tile along the vector sequence with complete coverage. These probes were hybridized to the gDNA, and streptavidin-coated beads were used to pull down fragments containing vector sequence. After washing the beads and digesting the RNA probes, the captured fragments were purified, and sequenced. Sequencing was performed using single end Illumina MiSeq, with an average read length of 250bp. Reads from sequencing were first mapped to the vector to filter out reads without vector information, and then mapped to the genome to filter out vector only reads. Split reads and pileup information was used to determine the integration junction.

PCR-Based Method

Integration site analysis for rDG_IgG was also performed using a nested PCR method. A diagram of this process is shown in Supplementary Figure S5. gDNA was extracted from H_1 cells using NucleoSpin Tissue Kit (Takara Clontech). The gDNA was then digested with three restriction enzymes: DraI, SspI, and HpaI. Digested DNA was further purified using NucleoSpin Gel and PCR Clean-Up Kit (Takara Clontech). After purification, GenomeWalker adapters (taken from the Lenti-X Integration Site Analysis Kit, Takara Clontech) were ligated to the fragmented gDNA. Two successive rounds of PCR were then performed to amplify gDNA fragments containing vector sequence using the Advantage 2 PCR Kit (Takara Clontech). In order to capture the different integration sites, multiple vector specific primers were used that spanned the product

integration vector. First, PCR was done using one of primary vector specific primers and Adapter Primer 1 (AP1), which binds to the ligated adapter. The PCR reaction was diluted, and then added to a second reaction using a corresponding nested vector primer and Adapter Primer 2 (AP2), found downstream of the original set used for amplification. Gel electrophoresis was used to visualize the nested PCR reaction, and resulting bands were extracted from the gel and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research). The purified PCR products were Sanger sequenced at the University of Minnesota Genomics Center (Saint Paul, MN).

GENOMIC DNA QUANTITATIVE PCR (QPCR)

Genomic DNA was isolated using NucleoSpin® Tissue Kit (Macherey-Nagel, Duren, Germany). The qPCR assay was conducted in triplicate using 20ng of genomic DNA in a 20 μL reaction volume on a CFX Connect Real-Time PCR Detection System (Bio-Rad, Hercules, CA) machine using SYBR Select Master Mix (Applied Biosystems, Foster City, CA). The cycle numbers obtained from the qPCR experiment were normalized to an unamplified region in the genome, β-actin, to account for differences in DNA loading. Data is reported as fold change relative to parent. The raw cycle number for β-actin was similar for all cell lines.

QRT-PCR FOR QUANTIFYING TRANSCRIPT LEVELS

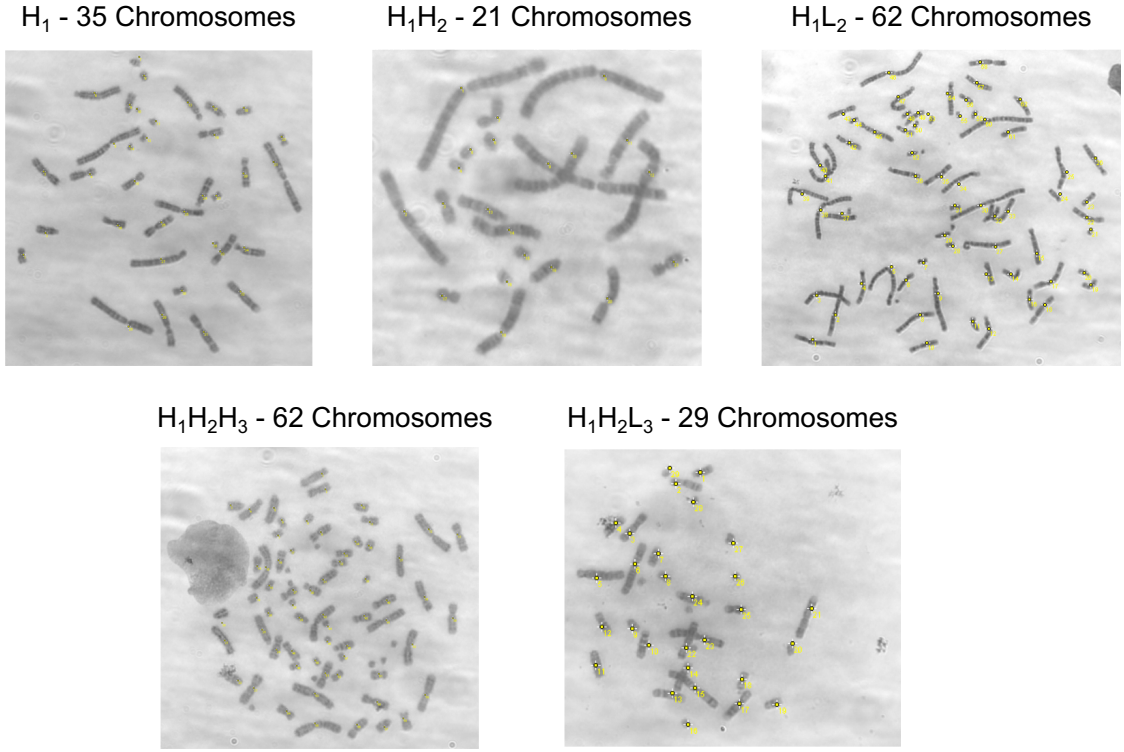
RNA was extracted from cells using RNeasy Mini Kit (Qiagen), and cDNA synthesis was done using SuperScript III First-Strand cDNA Synthesis SuperMix. qRT-PCR was performed using SYBR Select Master Mix (Applied Biosystems) on a CFX Connect Real-Time PCR Detection System (Bio-Rad, Hercules, CA). qRT-PCR data for the subclones was normalized to GAPDH, and then to their respective parents. Data is reported as fold change relative to parent.

SUPPLEMENTARY TABLES

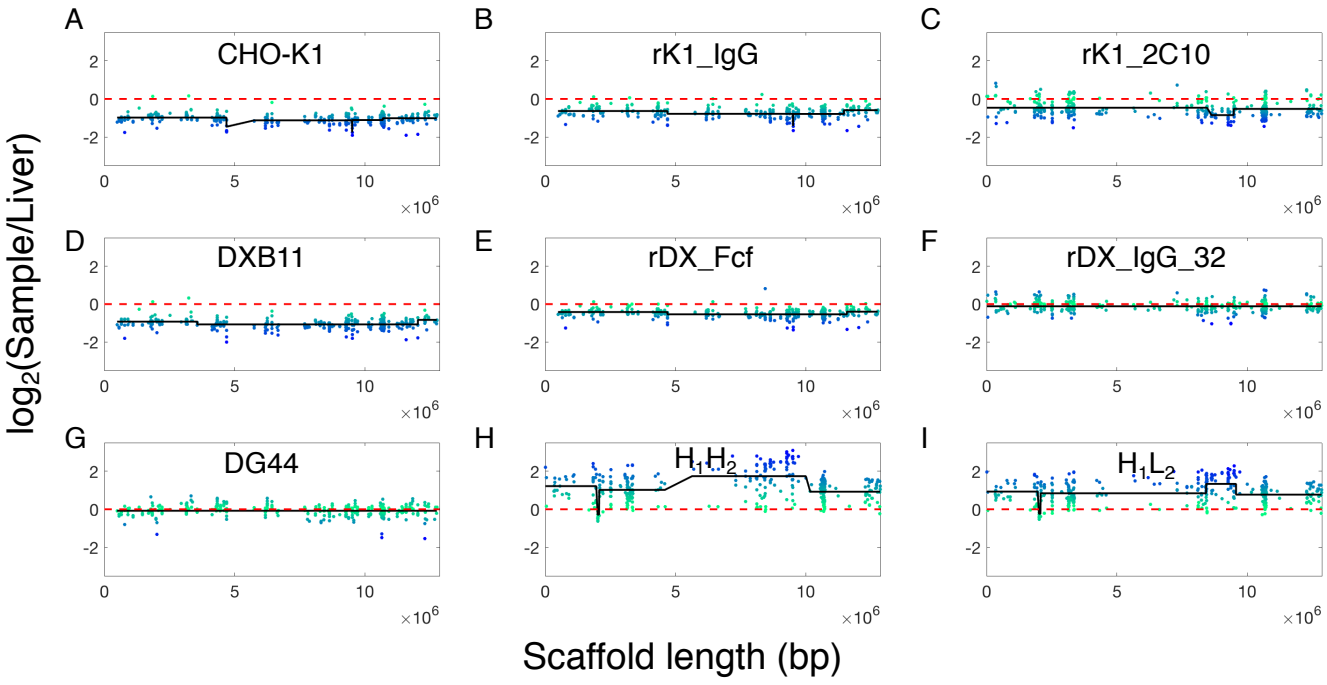
Supplementary Table S1. Description of integration sites in rDG_IgG and methods used for identification. Prime (‘) denotes other side of integration site.

Site #	Integration Location	Vector Junction	Region Amplified	Detection Method		
				PCR + Sanger	DELLY	Sequence Capture
1	Rc3h1, intron	CMV	Yes	✓	✗	✓
2	Vps13b, intron	**β**-Lac	Yes	✓	✓	✓
2’	Vps13b, intron	SV40 pA	Yes	✓	✓	✗
3	IGR	KanR	No	✗	✓	✓
3’	IGR	KanR/ SV40 pA	No	✗	✓	✗
4	Stag2, intron	**β**-Lac	No	✓	✓	✗
5	CYP3A31, intron	DHFR	No	✗	✓	✗
5’	CYP3A31, intron	DHFR	No	✗	✓	✗

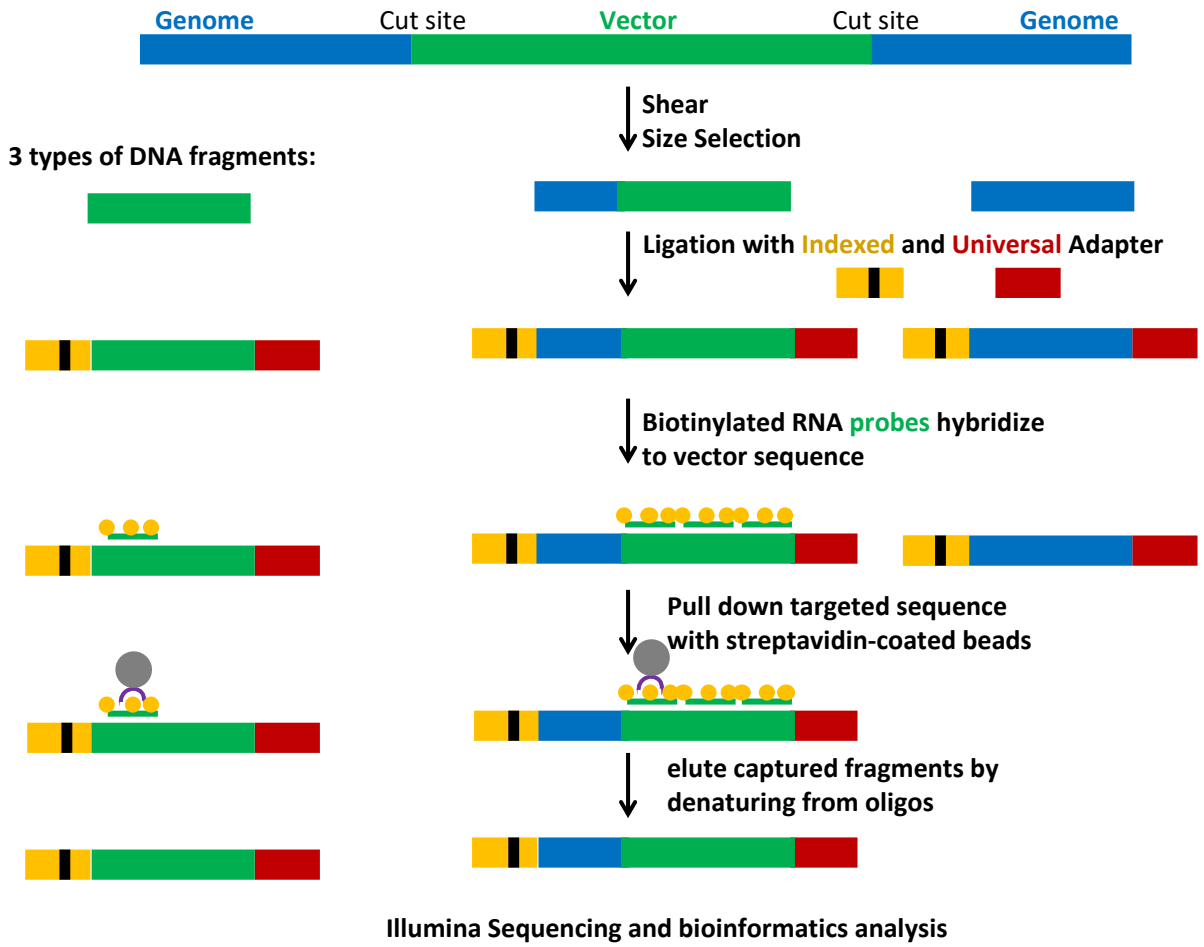
SUPPLEMENTARY FIGURES



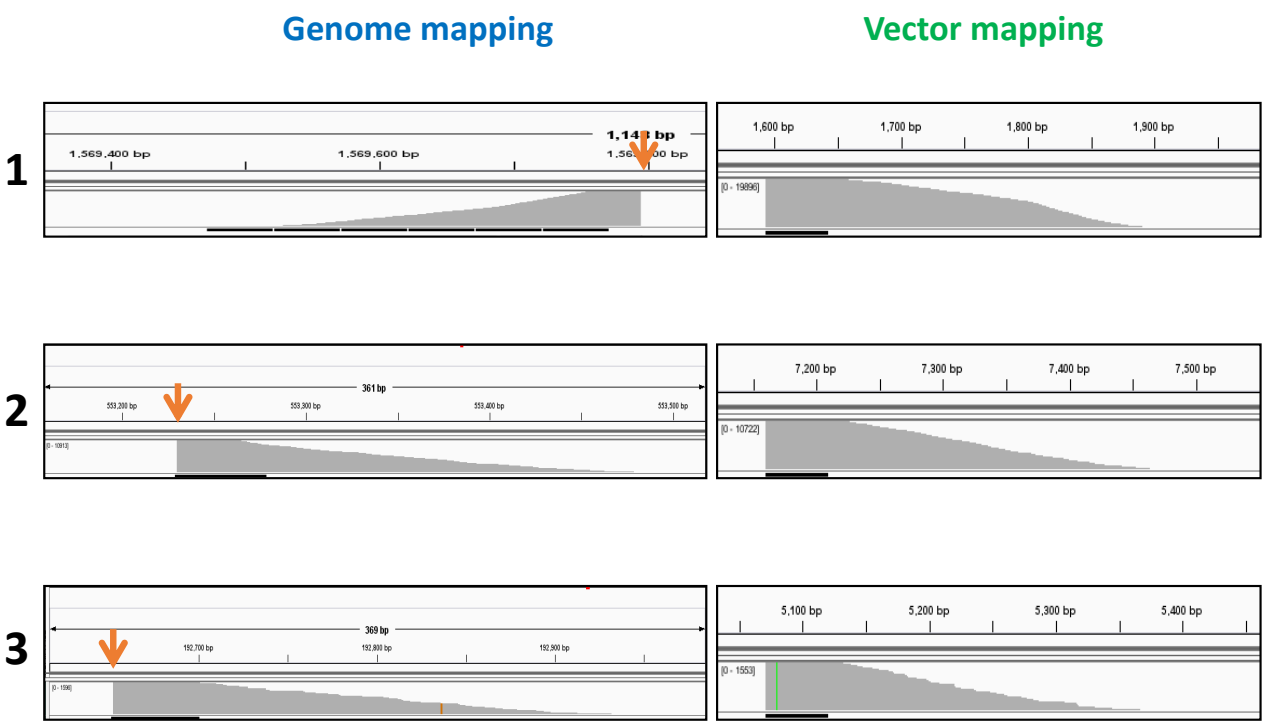
Supplementary Figure S1. Representative images from metaphase spreads. Chromosomes counted from image are marked in yellow.



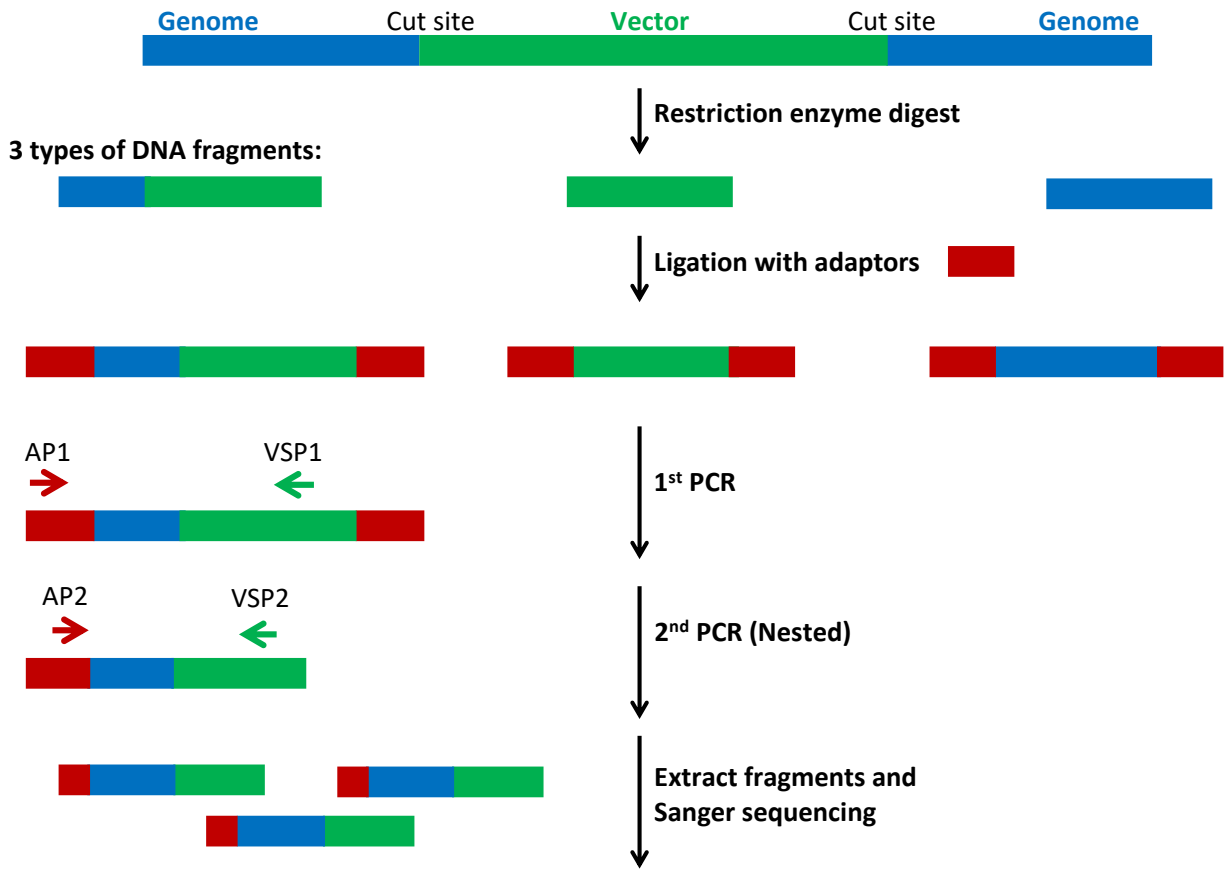
Supplementary Figure S2. Copy number variation in the genomic scaffold (shown in Figure 4) in cell lines from other lineages. **(A)** CHO-K1 and derived cell lines rK1_IgG **(B)** and rK1_2C10 **(C)**. **(D)** DXB11 originally derived from CHO-K1 and its derived cell lines rDX_Fcf **(E)** and rDX_32 **(F)**. **(G)** DG44 and its derived cell lines H_1H_2, H_1L_2. The green-blue dots represent the log-ratio intensities for each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy.



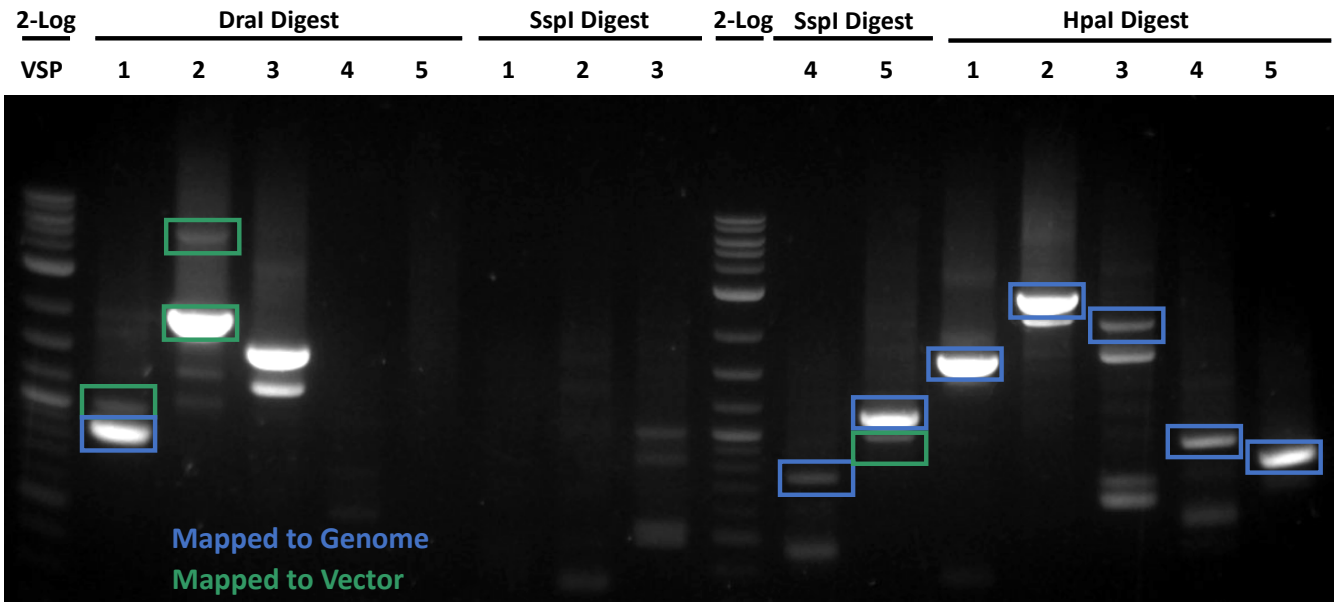
Supplementary Figure S3. Sequence capture based integration site analysis method. Genomic DNA was first sheared and size selection was used to isolate DNA in the 300-500bp range. Illumina sequencing adapters were then ligated, and biotinylated RNA probes and streptavidin beads were used to capture fragments containing vector sequence. Isolated fragments were sequenced using Illumina sequencing. See Supplementary Methods for a detailed description.



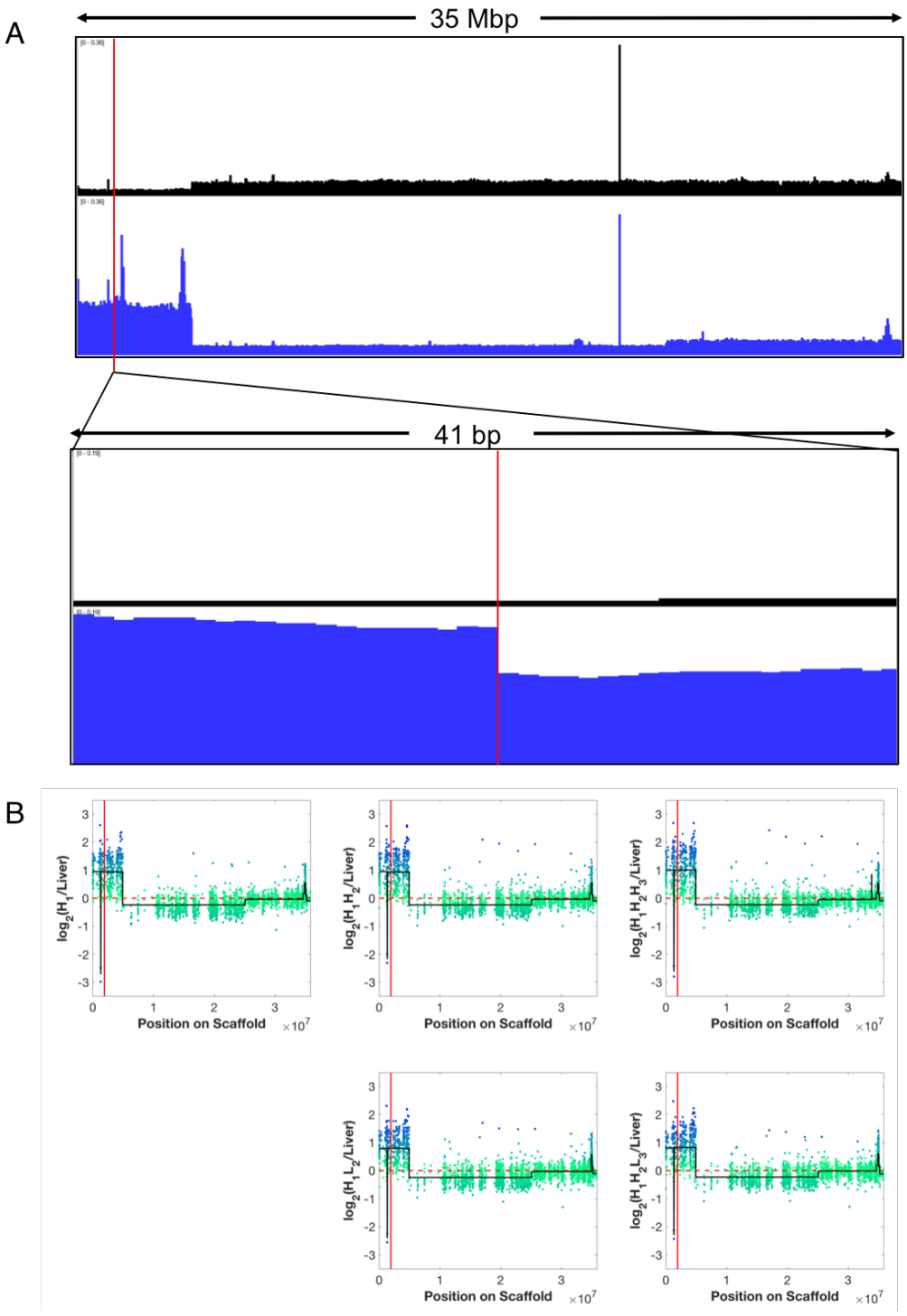
Supplementary Figure S4. Read pile up for sequence capture based integration site analysis. Orange arrows denote integration junctions.



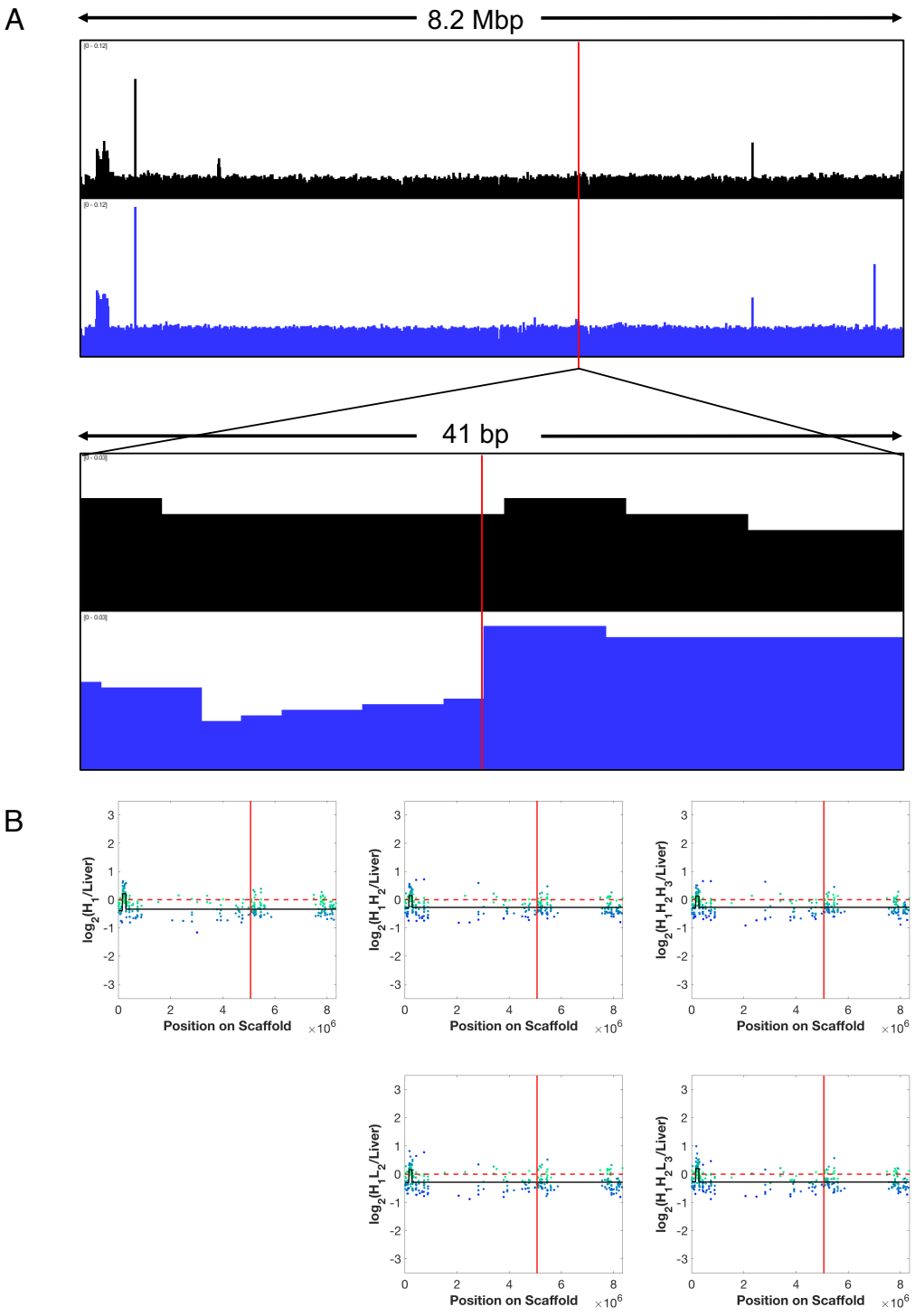
Supplementary Figure S5. Nested PCR based integration site analysis method. Genomic DNA was first fragmented using restriction enzymes and then ligated with an adaptor. Two rounds of PCR using a vector specific primer (VSP) and an adaptor specific primer (AP) are used to enrich for fragments containing both vector and genome sequences. Extracted fragments are then sequenced using Sanger sequencing. See Supplementary Methods for a detailed description.



Supplementary Figure S6. Gel electrophoresis of nested PCR for PCR-based integration site analysis. Extracted fragments are denoted with a box. (Blue: mapped to genome; Green: mapped to vector). Restriction enzyme library used to generate fragments is listed above lanes. VSP1-5 denote the vector specific primer used to amplify the fragments. (2-Log: 2 Log Ladder, NEB).



Supplementary Figure S7. (A) Normalized sequencing read pileup (Reads per Million) for DG44 and H_1 for integration site #2. The red vertical line denotes the integration junction. **(B)** CGH data for the five sub lines show no significant change among high and low producing clones. The green-blue dots represent the log-ratio intensities each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. The red vertical line denotes the integration junction.



Supplementary Figure S8. (A) Normalized sequencing read pileup (Reads per Million) for DG44 and H_1 for integration site #3. The red vertical line denotes the integration junction. **(B)** CGH data for the five sub lines show no significant change among high and low producing clones. The green-blue dots represent the log-ratio intensities each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. The red vertical line denotes the integration junction.

SUPPLEMENTARY REFERENCES

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics, 5*(4), 557-572. doi:10.1093/biostatistics/kxh008

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics, 23*(6), 657-663. doi:10.1093/bioinformatics/btl646

Vishwanathan, N., Bandyopadhyay, A., Fu, H. Y., Johnson, K. C., Springer, N. M., & Hu, W. S. (2017). A comparative genomic hybridization approach to study gene copy number variations among Chinese hamster cell lines. *Biotechnol Bioeng, 114*(8), 1903-1908. doi:10.1002/bit.26311