Long-read sequencing reveals reduced genomic heterogeneity in RHAMM-deficient metastases

Biology 4999E Honours Thesis

Student: Audrey Anna Goddard

Student ID: 251287880

Supervisor: Dr. Kathleen Hill

Advisor Panel: Dr. Robert Cumming & Dr. Anthony Percival-Smith

Mentor: Joseph Butler

Department of Biology, University of Western Ontario

April 2, 2026

Citation Style: Nature Genetics

Abstract

Genetic heterogeneity is a defining feature of breast cancer (BrCa), driving metastatic progression, therapeutic resistance, and poor patient outcomes across subtypes with distinct clinical behaviors. RHAMM (Receptor for Hyaluronan-Mediated Motility; HMMR), a regulator of centrosome function, ERK signaling, and genomic instability, is overexpressed in aggressive BrCa; however, its role in shaping clonal diversity during metastasis remains unclear. This study examined whether RHAMM influences genomic diversity in MMTV-PyMT lung metastases on a mixed FVB/N × C57BL/6 background, comparing wild-type (WT) and RHAMM-deficient (KO) cohorts. High-molecular-weight (HMW) DNA was extracted from bulk metastatic lung tissue and sequenced using PacBio HiFi long-read sequencing (mean coverage 48.38×; mean quality Q23.2). Reads were aligned to mm39 using pbmm2; variants were called with DeepVariant and phased with HiPhase. A multi-stage pipeline removed low-confidence, low-VAF, and strain-specific variants to isolate metastasis-associated calls. After stringent filtering, *Rhamm*$^{-/-}$ samples exhibited a 1.6-fold higher mutation burden than *Rhamm*$^{+/+}$ controls (mean 23,159 vs. 14,349 variants per sample; $p = 0.002$) but markedly reduced inter-animal variability. A total of 4,078 variants were exclusively shared across all *Rhamm*$^{-/-}$ mice, of which 82% had VAF ≥ 0.8 (mean VAF = 0.939), indicating clonal dominance. In contrast, only 269 variants were uniquely shared across all *Rhamm*$^{+/+}$ mice, with broader VAF distributions (34% with VAF ≥ 0.8; mean VAF = 0.695). The 4,078:269 split was significant (binomial test, $p < 2.2 \times 10^{-16}$). These findings support the hypothesis that RHAMM promotes genetic heterogeneity, while its loss drives clonal homogeneity and selection of a dominant metastatic genotype adapted to the lung microenvironment.

Keywords: RHAMM; HMMR; heterogeneity; breast cancer metastasis; PacBio HiFi long-read sequencing; MMTV-PyMT; clonal selection.

Supplementary data files are available in the GitHub repository at:
https://github.com/audreygoddard04/RHAMM-HiFi-lung-metastasis

Introduction

Metastasis is responsible for the majority of breast cancer mortality, with approximately 90% of cancer deaths attributable to metastatic disease rather than the primary tumor.[1] Among breast cancer subtypes, approximately 30% of Luminal B patients develop lung metastasis.[2] The ability of tumors to evolve genetically during progression and colonize distant organs is closely linked to genetic heterogeneity.[3] Genetic heterogeneity allows subclones within a tumor to acquire traits that enable survival in circulation, resistance to therapy, and adaptation to foreign microenvironments.[3,4] Understanding the molecular factors that regulate genetic heterogeneity is therefore critical for identifying vulnerabilities in metastatic disease.

RHAMM (Receptor for Hyaluronan-Mediated Motility; HMMR) is a hyaluronan-binding protein with both intracellular and extracellular functions.[5,6] Intracellular RHAMM localizes to centrosomes, mitotic spindles, and microtubule arrays, where it contributes to spindle stability and chromosome segregation during cell division.[5] Through interactions with hyaluronan and CD44 at the cell surface, RHAMM also influences ERK signaling pathways that regulate proliferation, motility, and cellular responses to stress.[7] Dysregulation of these processes can promote genomic instability and tumor progression.

Although RHAMM overexpression is consistently associated with poor prognosis in breast cancer, its role in generating or maintaining tumor genetic heterogeneity remains unclear. Previous studies have demonstrated that RHAMM influences mitotic integrity and cytoskeletal organization, suggesting a potential mechanism by which RHAMM activity could increase mutational diversity within tumors. However, few studies have directly examined genome-wide mutational patterns in metastatic tumors under conditions of RHAMM loss.[7] Prior work by Tolg et al. demonstrated that *Rhamm*$^{-/-}$ mice in the MMTV-PyMT background develop more numerous lung metastases than wild-type counterparts, establishing that RHAMM shapes the metastatic process and motivating direct examination of the mutational landscape of these lesions.[7,8]

The polyoma middle T antigen under mouse mammary tumour virus promoter (MMTV-PyMT) mouse model provides a well-characterized system for studying metastatic breast cancer.[8–10] This transgenic model recapitulates key features of human luminal-like breast cancer, including

rapid tumor progression and spontaneous lung metastasis. In this study, the model was maintained on a mixed FVB/N × C57BL/6 background and crossed with RHAMM mutant mice to generate RHAMM-deficient and wild-type cohorts. Because mixed genetic backgrounds introduce substantial germline variation, rigorous computational filtering is required to distinguish metastasis-associated variants from strain-specific polymorphisms.[10,11]

Long-read sequencing technologies offer advantages for detecting complex genomic variation and reconstructing clonal structure.[12] PacBio HiFi sequencing generates circular consensus reads that combine long molecule lengths with high base accuracy, enabling reliable identification of single-nucleotide variants and other genomic alterations across the genome.[13]

The central research question of this study is: Does RHAMM expression influence genome-wide genetic heterogeneity in lung metastases? The hypothesis tested is that if RHAMM expression promotes clonal diversification in metastases, then RHAMM loss is associated with increased clonal homogeneity and selection of lung-adapted genotypes.

Materials & Methods

Mouse Model

This project examined lung metastasis-associated genomic variation in two cohorts of MMTV-PyMT transgene mice. *Rhamm*$^{-/-}$ animals were generated by crossing C57BL/6 *Rhamm*$^{-/-}$ with FVB/N MMTV-PyMT mice, producing F1 hybrid cohorts on a mixed FVB/N × C57BL/6 background. Two cohorts were established: RHAMM wild-type (*Rhamm*$^{+/+}$; MMTV-PyMT, n=4) and RHAMM-deficient (*Rhamm*$^{-/-}$; MMTV-PyMT, n=4). The mixed background was analytically important because inherited strain-specific variants could contribute substantially to inter-animal variation; downstream filtering explicitly accounted for both FVB-associated and C57BL/6-associated background variants. Lung tissue containing metastatic lesions was collected for genomic analysis. Because samples consisted of bulk lung tissue rather than isolated metastatic nodules, the sequenced DNA contained a mixture of metastatic, stromal, immune, and non-metastatic lung epithelial cells. Metastasis-specific somatic variants identified in this study

therefore reflect their frequency across the entire cellular mixture rather than within the tumor clone alone.

Furthermore, it should be noted that the RHAMM mutation employed in this study is not a complete gene deletion; the allele harbours a frameshift in a critical 5-prime coding exon expected to severely disrupt protein function, but residual function from alternative start sites or splice variants cannot be excluded. The distinction between partial loss-of-function and null alleles was maintained in the interpretation of all genotype comparisons. In addition, prior literature indicates that some wild-type metastases may exhibit reduced RHAMM expression at the metastatic stage,[8] which was considered in the biological interpretation of cohort differences.

DNA Extraction

Lung tissue samples were homogenized prior to extraction to ensure that metastatic lesions were evenly represented in the DNA preparation, minimizing subsampling bias. Tissue input masses ranged from 30–40 mg per sample (Table 1). High-molecular-weight (HMW) genomic DNA was extracted using the Nanobind PanDNA kit (PacBio) according to the manufacturer's protocol. This method uses a nanostructured silica surface that selectively binds large DNA molecules under appropriate buffer conditions while preserving long fragment lengths. Homogenization was performed using a TissueRuptor II with a low-shear setting to minimize mechanical fragmentation of high-molecular-weight DNA. Extractions were carried out in Protein LoBind tubes throughout. Purified DNA was eluted in nuclease-free water. Each sample was eluted in 130 µL of nuclease-free water; total DNA mass was calculated as the product of NanoDrop concentration and elution volume (Table 1).

DNA Quality Control

DNA concentration and purity were assessed using NanoDrop spectrophotometry (A260/280 ratios). DNA integrity was evaluated using 0.8% agarose gel electrophoresis (SYBRGreen, 1× TAE buffer) alongside a λ DNA–Mono Cut Mix HMW ladder (New England Biolabs) and by TapeStation capillary electrophoresis at The Centre for Applied Genomics (TCAG), SickKids Hospital, Toronto. The TapeStation system separates DNA fragments by size through a narrow capillary under an applied electric field and generates a DNA Integrity Number

(DIN) reflecting the degree of fragmentation on a scale of 1–10. Acceptance criteria were: A260/280 ≥ 1.8 and DIN ≥ 9. All eight samples met these thresholds prior to library preparation (Table 1).

PacBio HiFi Sequencing

Sequencing libraries were prepared at TCAG using the SMRTbell Express Template Prep Kit 2.0, and 3.0 µg of HMW genomic DNA per sample was used as input. DNA ends were repaired and SMRTbell hairpin adapters were ligated to both ends to generate circular SMRTbell templates. Libraries were sequenced on the PacBio Sequel IIe platform.

During circular consensus sequencing (CCS), a single DNA polymerase traverses the circular SMRTbell template multiple times. The resulting subreads from each pass are computationally merged to generate a high-fidelity consensus read. PacBio HiFi reads typically exceed 10 kb in length with consensus accuracy approaching 99.9%, making them suitable for reliable genome-wide SNV detection, structural variant calling, and haplotype phasing.[12]

Laboratory Computational Working Environment

The laboratory server is located on site in a central data centre and provides 200TB of primary ZFS storage to client users, and 200TB of secondary ZFS storage for full incremental backups. Storage is accessible to all other systems associated with the laboratory, via NFS, over dual redundant 10G ethernet connectivity to campus. The laboratory compute server is also located in the Western University Central Datacenter and is designed to provide raw computational power via standard CPU and GPU resources. These resources consist of dual Intel GOLD 6542Y CPUs consisting of 48 cores, 96 threads, and an NVIDIA RTX 6000 Ada generation GPU. The GPU provides 48GB GDDR6 memory, 18k CUDA cores, 568 Tensor cores, 142 Raytrace cores. The compute server has 256 GB of main memory, 1TB of high-performance local storage and direct access to the storage from kh-server over dual redundant 10G ethernet. Three laboratory workstations have the following specifications: Linux-based lab workstation with Intel Core i9-14900k CPU, 64GB ram, 4TB internal storage, internal graphics. A fourth workstation has the following specifications: Windows 11-based lab workstation with Intel Core Ultra 5-235 CPU, 32GB ram, 1 TB internal storage, internal graphics.

Read Alignment and Sequencing Quality Control

HiFi reads were aligned to the mouse reference genome (GRCm39/mm39) using pbmm2, a PacBio-optimized wrapper for minimap2. Resulting BAM files were sorted and indexed using samtools (v1.19.2)[14]. Post-alignment quality metrics were assessed using NanoPlot, NanoComp, mosdepth, and MultiQC (v1.33).[15,16]

Variant Calling

Single nucleotide variants (SNVs) and small insertions or deletions (indels) were identified using DeepVariant (v1.6.0),[17] a deep-learning-based variant caller optimized for PacBio HiFi data. Variant phasing was performed using HiPhase (v1.6.0)[18] to group variants occurring on the same haplotype into phased blocks. The resulting phased VCF files served as input for all downstream filtering and cohort comparison steps.

Variant Filtering

Variant filtering was performed in multiple sequential stages to maximize specificity for metastasis-associated variants while removing germline background polymorphisms, low-quality calls, and strain-specific variants (Table 7). Prior to strain filtering, RefSeq-style chromosome identifiers were converted to chromosome number format to ensure compatibility with Sanger Institute FVB and B6 reference datasets.

Stage 1 - Low VAF filtering: Variants with VAF < 0.30 at approximately 48× depth were removed. This threshold was chosen to eliminate potential sequencing artifacts and extremely low-frequency variants unlikely to represent dominant clonal populations. At a mean depth of 48×, a variant allele requires a minimum of approximately 14 supporting reads to reach VAF 0.30, providing a conservative but statistically grounded confidence threshold for distinguishing real variants from sequencing noise.

Stage 2 - B6 strain background filtering: Variants present in a C57BL/6-derived reference VCF (12_L1 endoderm cohort) were subtracted using bcftools isec to remove inherited C57BL/6 germline polymorphisms.

Stage 3 - FVB/N strain background filtering: Variants present in a FVB/N strain SNP/indel VCF, generated by remapping the FVB reference genome to the C57BL/6 reference, were subsequently subtracted. Together, Stages 2 and 3 constituted a two-step strain filtering approach targeting both parental strain backgrounds present in the mixed FVB/N × C57BL/6 cohort.

Stage 4 - Stringent quality filtering: The following thresholds were applied to retain only high-confidence variant calls: QUAL ≥ 50 (Phred-scaled variant quality; corresponding to ≤ 0.001% error probability), GQ ≥ 30 (99.9% genotype confidence), DP ≥ 30 (minimum total read depth), and AD[ALT] ≥ 5 (minimum five reads supporting the alternate allele).

Stage 5 - Cohort merging and comparison: Filtered single-sample VCFs were merged within each cohort using bcftools merge. Variants present in all four mice of a given cohort were retained. Cohort-exclusive variants were then determined by bcftools isec comparison between cohorts, identifying variants present in all four KO mice but absent from all four WT mice, and vice versa. Variant allele frequency distributions across animals were then evaluated to assess clonal patterns within each cohort.

Statistical Analysis

Differences in per-sample variant burden between *Rhamm*$^{-/-}$ and *Rhamm*$^{+/+}$ cohorts were assessed using a two-sample t-test on the mean count of variants across KO and WT samples (n=4 per group). A p-value of < 0.05 was considered statistically significant.

The statistical significance of the 4,078 vs. 269 cohort-exclusive variant split was assessed using a binomial test. Under the null hypothesis that variants are equally likely to be exclusive to either cohort ($p = 0.5$), the probability of observing 4,078 or more KO-exclusive variants out of 4,347 total was calculated using binom.test() in R. This yielded $p < 2.2 \times 10^{-16}$, confirming that the observed distribution is highly inconsistent with random chance.

VAF distributions between KO-exclusive and WT-exclusive variant sets were compared using a two-sample Kolmogorov-Smirnov test (ks.test() in R), which assesses whether the two distributions differ significantly in shape without assuming normality. All statistical analyses were performed in R (v4.5.0).

Functional Annotation Analysis

To investigate the biological functions represented by cohort-exclusive variants, genes associated with WT-exclusive and KO-exclusive variant sets were analyzed separately using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[19]. Gene symbols were extracted from SnpEff-annotated VCF files (annotated_KO_exclusive_variants.vcf and annotated_WT_exclusive_variants.vcf) using bcftools query on the ANN field and subsequently filtered to retain only entries beginning with a standard alphabetic character to remove non-gene annotations.[20] This yielded 1,897 unique KO-exclusive gene symbols and 434 unique WT-exclusive gene symbols for submission to DAVID. Gene symbols were submitted as Mus musculus identifiers. Functional annotation clustering was applied to group related annotation terms into higher-order biological themes, drawing on Gene Ontology biological process, cellular component, and molecular function categories, along with UniProt keywords, UniProt sequence features, InterPro domains, SMART domains, and KEGG pathways. Default DAVID clustering parameters were used. Cluster enrichment scores were used as a summary measure of functional overrepresentation; representative terms from the highest-scoring clusters were selected for biological comparison between cohorts.

As an independent validation, gene lists were additionally analyzed using ClusterProfiler (v4.18.4)[21] in R with org.Mm.eg.db for gene ID mapping and Benjamini-Hochberg correction for multiple testing. GO Biological Process, GO Cellular Component, and KEGG pathway enrichment were performed for both cohorts. ClusterProfiler results are presented in Supplementary Figures and were concordant with the primary DAVID findings. Due to the low mapping rate of non-standard gene symbols (approximately 21% for KO and 17% for WT), DAVID results are presented as the primary analysis.

Results

DNA Extraction and Quality Control: Sequencing Preparation

High-molecular-weight (HMW) DNA was successfully extracted from lung metastatic tissue for all eight samples using the Nanobind PanDNA protocol. Gel electrophoresis confirmed the presence of HMW bands with minimal low-molecular-weight smearing (Figure 1A). DNA quality was consistently high across the cohort. NanoDrop A260/280 ratios ranged from 1.85 to 1.89, indicating minimal protein contamination. DNA yields from NanoDrop ranged from 150 to 222 ng/µL following extraction, with total calculated masses of 19,500–28,860 ng per sample (Table 1). All eight samples achieved a TapeStation DIN score of 9.9 (Figure 1B), consistent with highly intact genomic DNA suitable for PacBio HiFi library preparation. Genome equivalent estimates based on 3.0 µg input for library preparation ranged from 2.79×10^6 to 4.12×10^6 cell equivalents, confirming adequate representation of the cellular population at sequencing depth (Table 2).

Sequencing Quality

PacBio HiFi sequencing on the Sequel IIe platform generated high-quality long-read data across all eight samples (Table 2). The mean number of reads per sample was 9,220,868 (range: 8,607,484–9,857,604). Mean read length averaged 18,963 bp (range: 17,511–21,123 bp) and N50 values ranged from 17,915 to 22,063 bp, with a cohort mean of 19,580 bp. Mean read quality scores ranged from Q22.5 to Q24.1 (Figure 2), with an overall average of Q23.2, corresponding to a base accuracy of approximately 99.5%, consistent with expected PacBio HiFi performance and above the Q20 minimum typically required for high-confidence SNV calling.

Genome coverage was similarly robust across all samples (Figure 2; Table 2). Target coverage was 25–40× per sample; mean achieved coverage was 48.38×, exceeding target requirements. Mean depth ranged from 42.20× (sample 4_4) to 51.53× (sample 7_5). Breadth of coverage at ≥1× was uniformly 94% for all samples, as was breadth at ≥10×. Coverage at ≥30× ranged from 85% (sample 4_4) to 91% (samples 7_4, 7_5, 7_6, 14_8). Together, these results

indicate that all samples achieved sufficient read quality and depth for high-confidence variant detection and downstream cohort-level comparison.

Variant Detection & Filtering

Variant calling identified genome-wide SNVs and indels across all eight samples. Raw pre-filter variant counts for KO cohort samples were as follows: 42_1: 5,073,242; 7_4: 5,875,627; 7_5: 6,158,094; 7_6: 5,034,425. Raw pre-filtered variants for WT cohort consisted of 14_8: 4,866,026; 27_7: 5,692,639; 4_2: 4,407,499; 4_4: 4,176,869. The majority of these raw variants represent germline polymorphisms, strain-background SNPs, and low-confidence calls that were removed during the subsequent filtering pipeline. After quality filtering and removal of background strain variants, high-confidence variant sets were obtained for each sample and used for cohort-level comparison.

Variant filtering was performed in five sequential stages to remove germline, strain-specific, and low-confidence variants and to isolate high-confidence metastasis-associated calls (Figure 3). The multi-stage approach was necessitated by the mixed FVB/N × C57BL/6 genetic background, which introduced large numbers of inherited polymorphisms shared between all animals regardless of genotype. After low VAF filtering (Stage 1), and two-stage strain filtering targeting C57BL/6 and FVB/N (Stage 2 & Stage 3) background variants, variant counts were substantially reduced from the raw totals.[22] Stringent quality filtering (Stage 4: QUAL ≥ 50, GQ ≥ 30, DP ≥ 30, AD[ALT] ≥ 5) retained only high-confidence calls. Cohort merging retained variants shared across all four mice within each genotype group, and final bcftools isec comparison identified 4,078 KO-exclusive and 269 WT-exclusive variants (Stage 5).

RHAMM-deficient metastases show increased cohort-shared variants

Comparison of filtered variants across animals within each cohort revealed substantial differences between RHAMM-deficient and wild-type metastases. A total of 4,078 variants were shared across all four *Rhamm*$^{-/-}$ mice and absent from all four wild-type mice, whereas only 269 variants were uniquely shared across all four *Rhamm*$^{+/+}$ mice (Figure 5). The RHAMM-deficient cohort therefore harboured approximately 15-fold more cohort-exclusive shared variants than the

wild-type cohort. This distribution was highly statistically significant (binomial test, $p < 2.2 \times 10^{-16}$).

At the individual sample level, mean per-sample variant burden was higher in the KO cohort than in the WT cohort (mean 23,159 vs. 14,349 variants per sample; $p = 0.002$, two-sample t-test) (Figure 4). Despite this higher overall mutation load, *Rhamm*$^{-/-}$ samples showed markedly reduced inter-animal variability and greater convergence toward a common variant set. Importantly, a 1.6-fold increase in overall mutation burden is insufficient to account for a 15-fold increase in cohort-exclusive shared variants, indicating that the observed homogeneity reflects a biological pattern of clonal selection. These data indicate that RHAMM loss is associated with increased clonal homogeneity at the cohort level; more variants are shared across all four animals, rather than simply increased total mutational burden per mouse.

Variant Allele Frequency Patterns

KO-exclusive variants were strongly enriched for high variant allele frequencies (Figure 6), consistent with clonal dominance. Among the 4,078 KO-exclusive shared variants, 82% had VAF ≥ 0.8, with a mean VAF of 0.939. This distribution indicates that the vast majority of KO-exclusive variants are carried by nearly all cells in the sequenced tissue, consistent with strong clonal selection for these alleles during metastatic colonization. In contrast, WT-exclusive shared variants displayed broader VAF distributions, and only 34% had VAF ≥ 0.8, with a mean VAF of 0.695, with the remaining variants occupying lower frequency bins indicative of subclonal populations. The VAF distributions of KO-exclusive and WT-exclusive variants were significantly different (Kolmogorov-Smirnov test, $D = 0.61$, $p = 1.22 \times 10^{-79}$). These findings further support the interpretation that RHAMM-deficient metastases are genetically more homogeneous and dominated by major clones. It should be noted that VAF values derived from bulk tissue sequencing reflect variant frequency across all cell types present, including stromal and immune cells.

Functional annotation clustering of WT- and KO-exclusive genes

Functional annotation clustering of cohort-exclusive genes revealed distinct enrichment patterns between RHAMM wild-type and RHAMM-deficient metastases (Figure 7; Tables 4, 5).

WT-exclusive genes were enriched for categories related to neuronal projections, cadherin-mediated adhesion, adherens junctions, and cell morphogenesis. The top clusters corresponded to synaptic and projection structures (enrichment score = 2.56) and adhesion and morphogenesis pathways (enrichment score = 2.16). The synaptic and neurological terms in Cluster 1 (glutamatergic synapse, dendrite, axon) are likely attributable to residual strain background variation from the FVB/N × C57BL/6 cross rather than metastasis-specific biology; both parental strains harbour variants in neurologically relevant genes that may not have been completely removed by the strain filtering steps. The cadherin-mediated adhesion and morphogenesis terms in Cluster 2 are more biologically interpretable in the context of metastatic tumor cells.

In contrast, KO-exclusive genes were most strongly enriched for low-complexity and intrinsically disordered protein regions (enrichment score = 3.55), followed by postsynaptic and synaptic membrane terms, membrane glycoproteins, RNA silencing and P-body functions, and RasGEF and Wnt signaling pathways. The enrichment of intrinsically disordered protein regions, flexible molecular hubs that mediate protein-protein interactions and regulate signaling network topology, in KO-exclusive variants suggests that the dominant clone has accumulated mutations capable of broadly rewiring cellular signaling. P-body and RNA silencing enrichment indicates altered post-transcriptional gene regulation.[23] The RasGEF and Wnt signaling enrichment is particularly relevant to RHAMM biology: RHAMM normally scaffolds ERK signaling through hyaluronan and CD44 interactions, and clones acquiring activating mutations in the RasGEF/RAS-ERK axis may carry a selective advantage in the context of RHAMM deficiency.[24]

These results suggest that WT-exclusive variants are more associated with structural, adhesion, and morphogenesis-related gene programs, whereas KO-exclusive variants are associated with membrane signaling and post-transcriptional regulatory pathways. ClusterProfiler GO Biological Process enrichment analysis further supported mitochondrial metabolism as a prominent theme among KO-exclusive genes (Figure 8). These functional interpretations represent predicted pathway associations based on gene-level annotation and should be considered exploratory; validation by orthogonal methods such as RNA sequencing or targeted functional assays would be required to confirm pathway involvement and establish causality.

Discussion

This study used PacBio HiFi long-read sequencing to characterize genome-wide genetic heterogeneity in lung metastases from MMTV-PyMT mice with and without RHAMM. The central finding is that *Rhamm*$^{-/-}$ lung metastases harboured approximately 1.6-fold higher mutation burden (23,159 vs 14,349) and 15-fold more cohort-exclusive shared variants than *Rhamm*$^{+/+}$ controls (4,078 vs. 269), while simultaneously displaying markedly reduced inter-animal variability and high variant allele frequencies indicative of clonal dominance. These results support the hypothesis that RHAMM loss is associated with clonal selection toward a dominant genotype during metastatic colonization of the lung.

The observation that KO metastases have both higher overall mutation burden and greater clonal homogeneity may appear paradoxical, but these findings are reconcilable. A higher total number of variants per sample does not indicate greater intratumoral heterogeneity if those variants belong to a dominant clone present across all cells in the tissue. The mean VAF of 0.939 among KO-exclusive variants indicates that the additional variants in KO samples are present in nearly all cells, consistent with early clonal dominance rather than ongoing subclonal diversification. By contrast, the broader VAF distribution in WT-exclusive variants (mean VAF 0.695; only 34% at VAF ≥ 0.8) reflects multiple subclonal populations at varying frequencies; the hallmark of an actively diversifying tumor.[22] Critically, a 1.6-fold increase in overall mutation burden is insufficient to account for a 15-fold increase in cohort-shared exclusive variants, and the observed split between KO- and WT-exclusive variants was highly statistically significant (binomial test, $p < 2.2 \times 10^{-16}$), confirming that the pattern of clonal homogeneity in KO metastases reflects a genuine biological difference.

The observed pattern of reduced heterogeneity in RHAMM-deficient metastases is interpretable through RHAMM's established roles in centrosome function and mitotic fidelity. In wild-type tumors, RHAMM contributes to spindle pole stability and proper chromosomal segregation during cell division.[4] Dysregulation of these functions can promote chromosomal instability (CIN),[25] driving the ongoing accumulation of copy-number alterations, structural variants, and point mutations during successive cell divisions, continuously expanding the mutational landscape and maintaining a heterogeneous population of genetically distinct clones.

In RHAMM-deficient tumors, the data are more consistent with a model of early mutation accumulation followed by strong clonal selection during metastatic colonization. Although KO metastases exhibit higher overall mutation burden, the majority of these variants occur at high variant allele frequency and are shared across animals, indicating early acquisition followed by clonal expansion. Under this model, one or a small number of clones best adapted to the pulmonary microenvironment undergo rapid expansion and outcompete other lineages, resulting in reduced heterogeneity at the endpoint. The high VAF observed in KO-exclusive variants supports this interpretation, as these variants appear to have undergone positive selection. Consistent with prior work by Tolg et al., which showed increased metastatic burden, the selective advantage of the dominant KO clone is more likely driven by phenotypic adaptation, such as enhanced colonization efficiency or microenvironmental fitness, rather than an increased mutation rate.

RHAMM also modulates ERK signaling through its interactions with hyaluronan and CD44.[6] In wild-type tumors, RHAMM-dependent ERK activation may promote proliferation and survival across a diverse range of subclonal backgrounds, enabling the maintenance of a heterogeneous cellular community.[26] In RHAMM-deficient tumors, reduced ERK scaffolding activity may confer a survival advantage selectively to clones that have acquired compensatory activating mutations in the RAS-ERK pathway or that express alternative proliferative signals, consistent with the RasGEF and Wnt signaling enrichment observed in KO-exclusive gene sets by DAVID functional annotation. This proposed compensatory mechanism, wherein loss of RHAMM-mediated ERK scaffolding creates selective pressure for clones that independently activate the same downstream pathway, represents a testable hypothesis for future RNA sequencing or targeted inhibitor studies.

The functional annotation analysis revealed distinct biological programs associated with WT- versus KO-exclusive variant sets. WT-exclusive genes were enriched for cadherin-mediated adhesion, adherens junctions, and cell morphogenesis; programs consistent with the ongoing epithelial plasticity and structural remodeling that characterize heterogeneous, actively evolving tumors. This is consistent with a model in which WT metastases maintain ongoing clonal diversification, with different clones exploring slightly different adhesion and motility strategies. KO-exclusive genes were enriched for membrane glycoproteins, postsynaptic signaling terms,

RNA silencing and P-body functions, and RasGEF and Wnt signaling pathways.[27] The enrichment of post-transcriptional regulatory functions, RNA silencing and P-bodies, in KO-exclusive genes is particularly intriguing: P-bodies are cytoplasmic condensates that regulate mRNA stability and translational repression, and their functional alteration could represent a mechanism by which the dominant KO clone modulates its gene expression landscape independently of new mutations.[23] These findings, while exploratory, suggest that RHAMM status may influence not only the extent of genetic convergence but also the functional classes of genes preferentially altered during metastatic evolution.

The enrichment of intrinsically disordered protein regions as the top DAVID cluster in KO-exclusive genes (enrichment score 3.55) deserves comment. Intrinsically disordered proteins are flexible molecular hubs that mediate protein-protein interactions across a wide range of partners and are disproportionately represented among oncoproteins and transcriptional regulators.[28] Mutations in disordered regions are often tolerated or beneficial in the context of signaling network rewiring, as they can alter interaction specificity without destroying protein function. Their accumulation in the dominant KO clone may therefore reflect broader rewiring of cellular signaling networks contributing to clone fitness in the lung microenvironment.

Pilot SNP array data from the Hill laboratory provided preliminary evidence that *Rhamm*$^{-/-}$ MMTV-PyMT lung metastases exhibit reduced genomic heterogeneity compared to wild-type controls, and prior work by Tolg et al. established that RHAMM regulates metastatic outcomes in this model through STING-dependent DNA damage sensing and interferon/STAT1 pro-apoptotic signaling. To our knowledge, the present study is the first to apply genome-wide long-read sequencing to characterize metastatic genetic heterogeneity in RHAMM-deficient tumors, confirming and substantially extending those preliminary observations across a broader set of variant classes, including phased SNVs and indels. Collectively, these findings demonstrate that RHAMM status influences not only the abundance and survival of metastatic populations, but also their underlying genetic diversity.

The observation that RHAMM loss is associated with clonal selection is also consistent with the broader cancer biology literature on metastatic bottlenecks.[29] Metastatic colonization of a distant organ requires survival in circulation, extravasation, adaptation to a novel

microenvironment, and establishment of a vascular niche, a series of selective pressures expected to dramatically narrow clonal diversity relative to the primary tumor.[3] In wild-type tumors, RHAMM-mediated genomic instability may continuously generate new clonal variants even after extravasation, maintaining diversity despite this bottleneck. In RHAMM-deficient tumors, reduced new variant acquisition may allow the bottleneck effect to fully manifest, resulting in metastatic lesions dominated by the clone best adapted to survive colonization. This model predicts that primary tumors from *Rhamm*$^{-/-}$ mice would show greater heterogeneity than matched lung metastases, a prediction that future primary tumor sequencing could directly test.

The use of PacBio HiFi long-read sequencing distinguishes this study from prior genomic analyses of tumor heterogeneity, which have predominantly relied on short-read Illumina sequencing.[30] Long reads provide several advantages: they span repetitive genomic regions inaccessible to short reads, enabling variant detection in gene families and low-complexity regions systematically missed by short-read approaches; haplotype phasing via HiPhase enables direct assessment of whether variants occur on the same or different chromosomal copies; and higher single-molecule accuracy through circular consensus sequencing reduces the false-positive rate for low-VAF variants. These properties make PacBio HiFi particularly well-suited to detecting clonal architecture in bulk tissue.

Separately, notable limitations must be considered when interpreting these findings. The most important limitation is the use of bulk lung tissue rather than isolated metastatic nodules. The sequenced DNA contains a mixture of metastatic tumor cells, lung stromal cells, immune infiltrates, vascular endothelium, and normal lung epithelium. Variants identified in this study cannot be assigned to specific cell populations. Metastasis-specific somatic variants may be diluted by the non-tumorigenic cell fraction, and the VAF observed for a given variant reflects its frequency across the entire cellular mixture rather than within the tumor clone alone. However, this limitation applies equally to both cohorts and is therefore unlikely to account for the directional difference in VAF distributions observed between KO and WT. Second, despite two-stage strain filtering targeting FVB/N and C57BL/6 background variants, residual strain-specific polymorphisms may remain, particularly at positions not represented in the FVB reference VCF or 12_L1 cohort dataset. However, as both cohorts share the same mixed FVB × B6 background, residual strain variants would be expected to affect both cohorts equally and would not

preferentially inflate KO-exclusive counts. The cohort-level recurrence filter, requiring a variant to be present in all four mice of one cohort and absent from all four of the other, provides a second layer of specificity that makes systematic strain background contamination unlikely. Third, each cohort contains four biological replicates (n=4 per group). This sample size limits statistical power for formal hypothesis testing and makes it difficult to distinguish biologically reproducible from stochastically variable findings. The cohort-level variant comparison approach (requiring a variant to be present in all four mice) mitigates some of this concern by enforcing stringent reproducibility, but the overall findings should be considered exploratory and hypothesis-generating pending validation in larger cohorts. Fourth, the RHAMM mutation used in this study is not a complete gene deletion, and the extent to which residual RHAMM function may influence the observed phenotype is unknown. Additionally, prior reports indicate that RHAMM expression may be reduced in wild-type metastases relative to the primary tumor,[12] which could attenuate the effective difference between cohorts at the metastatic stage. Finally, the DAVID and ClusterProfiler analyses are exploratory and subject to the well-documented limitations of gene set enrichment analysis, including sensitivity to gene list size, choice of background, and the presence of non-standard gene symbols that cannot be mapped to annotation databases.

Several extensions of this work would substantially strengthen its conclusions. Targeted validation of key cohort-exclusive variants by droplet digital PCR (ddPCR) in an expanded cohort (8–10 animals per group) would confirm allele frequencies by an orthogonal method and provide independent evidence for clonal dominance. Primary tumor sequencing would establish a temporal framework for the variant landscape, enabling determination of which KO- or WT-exclusive variants arose at the primary site versus during metastatic colonization, and directly testing the prediction that primary tumors show greater heterogeneity than matched metastases. Structural variant analysis using pbsv on the existing HiFi data would expand mutational characterization beyond SNVs and indels to include larger genomic rearrangements, which may also differ between cohorts. Laser capture microdissection of metastatic nodules prior to DNA extraction, or single-cell whole genome sequencing, would eliminate the bulk tissue limitation and enable direct clonal reconstruction at single-cell resolution. Single-cell RNA sequencing would provide clonal transcriptome profiles for direct assessment of intratumoral heterogeneity and identification of cell populations harboring cohort-exclusive variants. DNA methylation

analysis and mitochondrial genome assembly from the existing PacBio HiFi data represent additional analyses that would provide a more complete picture of epigenetic and metabolic heterogeneity.

Conclusion

This study demonstrates that RHAMM expression is associated with increased intrametastatic genetic heterogeneity in MMTV-PyMT lung metastases. Using PacBio HiFi long-read sequencing and a stringent multi-stage variant filtering pipeline, we show that *Rhamm*$^{-/-}$ metastases exhibit a dramatically higher number of cohort-shared exclusive variants (4,078 vs. 269 in wild-type), with high variant allele frequencies (mean VAF = 0.939) consistent with clonal dominance across all four knockout animals. These patterns are consistent with clonal selection in RHAMM-deficient tumors; the emergence of a dominant metastatic genotype that is shared across all animals in the cohort.

The functional annotation analysis further suggests that the genes affected by KO-exclusive variants are associated with distinct biological programs relative to WT-exclusive genes, with KO-exclusive variants predicted to be enriched for membrane signaling and post-transcriptional regulatory pathways; these associations are exploratory and require validation by orthogonal approaches. Together, these findings suggest that RHAMM contributes to the generation and maintenance of genetic diversity during metastatic colonization, and that its loss constrains this diversification process. Clonal homogeneity in RHAMM-deficient metastases may represent a therapeutic vulnerability, as tumors dominated by a single genotype could be more susceptible to targeted agents that exploit the shared dominant clone and less likely to generate resistant subclones.

These findings are exploratory given the bulk sequencing approach and small cohort size, and validation in larger cohorts using single-cell and orthogonal methods will be important next steps. Nonetheless, this work establishes a genome-wide long-read sequencing framework for characterizing metastatic clonal architecture in the MMTV-PyMT model and identifies RHAMM status as a determinant of the extent of clonal selection during lung colonization.

Acknowledgements

I would like to thank my advisory committee members, Dr. Robert Cumming and Dr. Anthony Percival-Smith, for their guidance, feedback, and thoughtful questions. Their interest in my project gave me a great deal of motivation and their multidisciplinary knowledge gave me further context. I am also grateful to Dr. Eva Turley, the pioneer of RHAMM, and Dr. Cornelia Tolg, who played a massive role in the wet lab work of my project. They both showed a great deal of generosity, and I am very grateful for their expertise and for their willingness to collaborate. Thank you, Joseph Butler, for the endless kindness, support, and advice. His help contributed greatly to the facilitation of navigating many aspects of this project. I would like to express my deepest gratitude to my supervisor, Dr. Kathleen Hill. Her mentorship, patience, and incredible sense of humor were fundamental to the completion of this thesis. Beyond her scientific guidance, she has a level of kindness that has had a lasting impact on me. I am deeply thankful for the opportunity to learn from her and for the example she sets as both a scientist and a person. Finally, I would also like to acknowledge the financial support that made this work possible. Funding was provided by the Department of Biology, the Faculty of Science, NSERC, and a generous anonymous donor.

References

1. Seyfried, T. N. & Huysentruyt, L. C. On the Origin of Cancer Metastasis. *Crit Rev Oncog* **18**, 43–73 (2013).
2. Fan, J.-H. *et al.* Molecular subtypes predict the preferential site of distant metastasis in advanced breast cancer: a nationwide retrospective study. *Front Oncol* **13**, 978985 (2023).
3. Lüönd, F., Tiede, S. & Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br J Cancer* **125**, 164–175 (2021).
4. Lv, D. *et al.* Exploration of the clonal evolution and construction of the tumor clonal evolution rate as a prognostic indicator in metastatic breast cancer. *BMC Med* **23**, 122 (2025).
5. Misra, S., Hascall, V. C., Markwald, R. R. & Ghatak, S. Interactions between Hyaluronan and Its Receptors (CD44, RHAMM) Regulate the Activities of Inflammation and Cancer. *Front Immunol* **6**, 201 (2015).
6. Turley, E. A., Noble, P. W. & Bourguignon, L. Y. W. Signaling Properties of Hyaluronan Receptors. *Journal of Biological Chemistry* **277**, 4589–4592 (2002).
7. Tolg, C. *et al.* RHAMM Promotes Interphase Microtubule Instability and Mitotic Spindle Integrity through MEK1/ERK1/2 Activity *. *Journal of Biological Chemistry* **285**, 26461–26474 (2010).
8. Tolg, C. *et al.* RHAMM regulates MMTV-PyMT-induced lung metastasis by connecting STING-dependent DNA damage sensing to interferon/STAT1 pro-apoptosis signaling. *Breast Cancer Res* **25**, 74 (2023).
9. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol Cell Biol* **12**, 954–961 (1992).

10. Attalla, S., Taifour, T., Bui, T. & Muller, W. Insights from transgenic mouse models of PyMT-induced breast cancer: recapitulating human breast cancer progression in vivo. *Oncogene* **40**, 475–491 (2021).

11. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155–1162 (2019).

12. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).

13. HiFi Reads - Highly accurate long-read sequencing. *PacBio* https://www.pacb.com/technology/hifi-sequencing/.

14. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

15. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

16. Mosdepth: quick coverage calculation for genomes and exomes - PubMed. https://pubmed.ncbi.nlm.nih.gov/29096012/.

17. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).

18. Holt, J. M. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).

19. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216–W221 (2022).

20. The Ensembl Variant Effect Predictor - PubMed. https://pubmed.ncbi.nlm.nih.gov/27268795/.

21. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

22. Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e39 (2021).

23. Parker, R. & Sheth, U. P bodies and the control of mRNA translation and degradation. *Mol Cell* **25**, 635–646 (2007).

24. Santarpia, L., Lippman, S. M. & El-Naggar, A. K. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opin Ther Targets* **16**, 103–119 (2012).

25. Thompson, S. L. & Compton, D. A. Chromosome missegregation in human cells arises through specific types of kinetochore-microtubule attachment errors. *Proc Natl Acad Sci U S A* **108**, 17974–17978 (2011).

26. Mebratu, Y. & Tesfaigzi, Y. How ERK1/2 Activation Controls Cell Proliferation and Cell Death Is Subcellular Localization the Answer? *Cell Cycle* **8**, 1168–1175 (2009).

27. Clevers, H. & Nusse, R. Wnt/β-catenin signaling and disease. *Cell* **149**, 1192–1205 (2012).

28. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology* **21**, 432–440 (2011).

29. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352**, 169–175 (2016).

30. Biada, I., Santacreu, M. A., González-Recio, O. & Ibáñez-Escriche, N. Comparative analysis of Illumina, PacBio, and nanopore for 16S rRNA gene sequencing of rabbit's gut microbiota. *Front. Microbiomes* **4**, (2025).

Figures

Figure 1. HMW DNA Isolation Quality Assessment

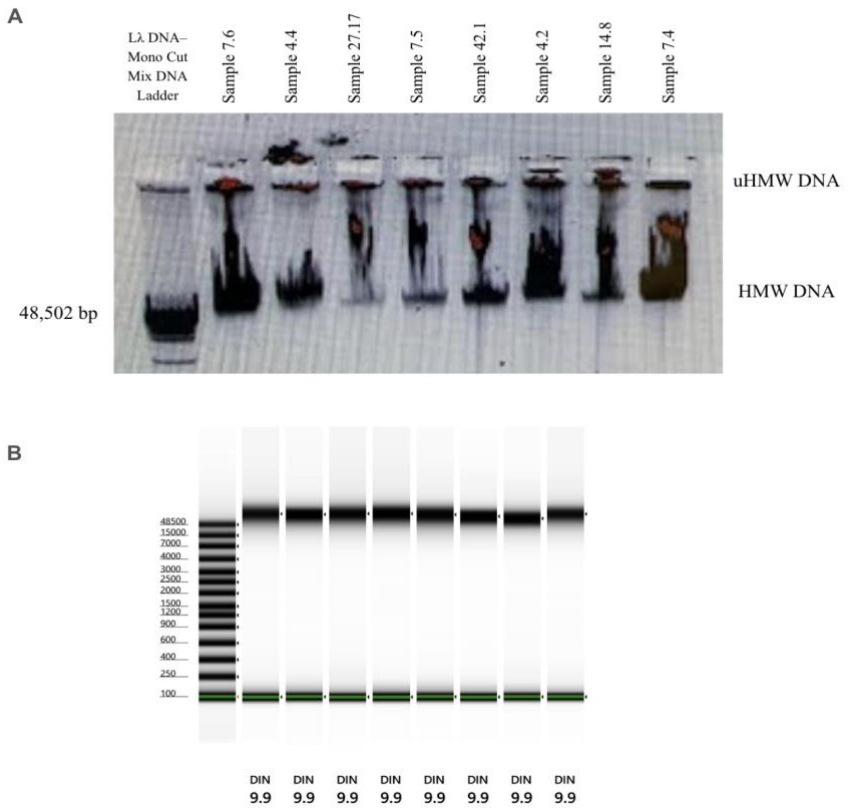

Figure 1. DNA Extraction and Quality Control. (A) Agarose gel electrophoresis (0.8%, SYBRGreen, 1× TAE) of all eight samples alongside a λ DNA–Mono Cut Mix HMW ladder (New England Biolabs; top band: 48,502 bp). The predominant high-molecular-weight band above the 48,502 bp marker confirms that the extracted genomic DNA is largely intact. The faint material at the top of each lane (uHMW DNA) indicates the presence of ultra-high-molecular-weight fragments. Minimal low-molecular-weight smearing is observed across all samples. (B) TapeStation capillary electrophoresis profiles for all eight samples. DNA Integrity Numbers (DIN) were uniformly 9.9 across all samples, confirming highly intact, non-degraded genomic DNA. All samples met acceptance criteria (A260/280 ≥ 1.8; DIN ≥ 9) prior to library preparation. Sample labels correspond to mouse IDs listed in Table 1. Green labels = *Rhamm*$^{-/-}$ (KO); white labels = *Rhamm*$^{+/+}$ (WT).

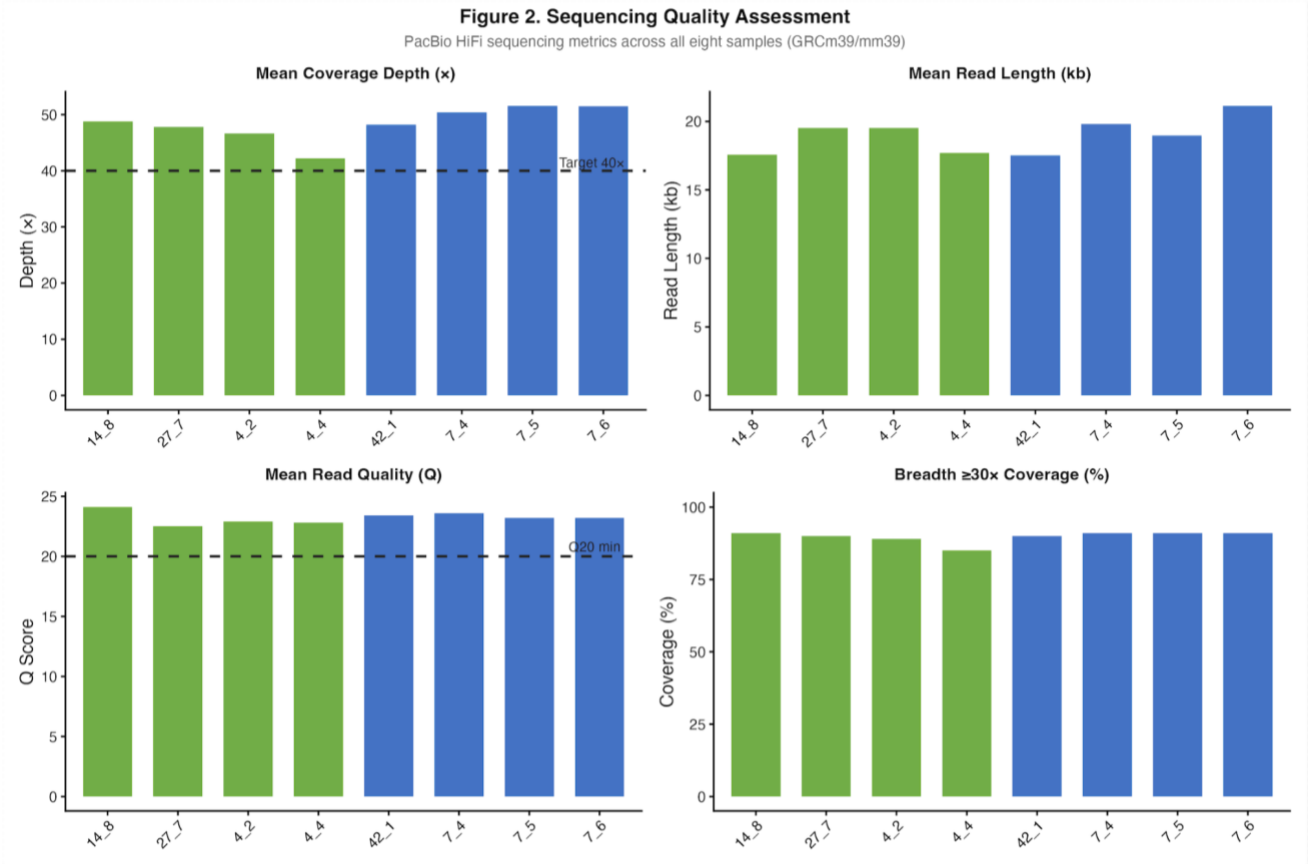

Figure 2. Sequencing Quality Assessment. Bar charts summarizing per-sample sequencing performance following circular consensus sequencing (CCS) generation and alignment to GRCm39/mm39. **(top left)** Mean coverage depth (×); the dashed line indicates the 40× target. All samples exceeded the 25–40× target range (mean 48.38×). **(top right)** Mean read length (kb); values ranged from 17,511 to 21,123 bp. **(bottom left)** Mean read quality (Q score); the dashed line indicates the Q20 minimum threshold for high-confidence SNV calling. All samples exceeded Q20 (mean Q23.2, corresponding to ~99.5% base accuracy). **(bottom right)** Breadth of coverage at ≥30× (%); all samples achieved ≥85% breadth at ≥30×. Green bars = *Rhamm*$^{-/-}$ (KO); blue bars = *Rhamm*$^{+/+}$ (WT). Full per-sample metrics are reported in Tables 4 and 5.

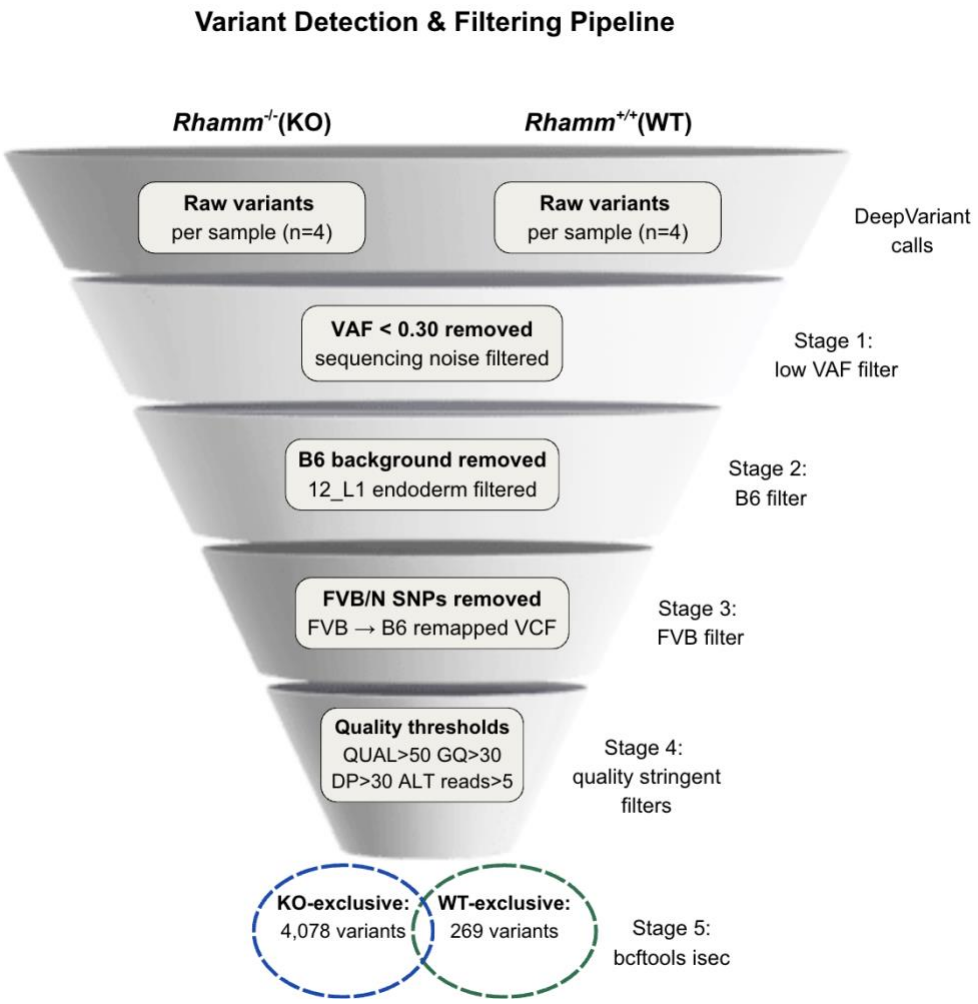

Figure 3. Variant Detection & Filtering Pipeline. Schematic funnel diagram illustrating the sequential variant filtering approach applied to both cohorts. Raw variants per sample were called using DeepVariant on PacBio HiFi reads. Stage 1: variants with VAF < 0.30 were removed to eliminate sequencing noise. Stage 2: C57BL/6 strain background variants were subtracted using bcftools isec against the 12_L1 endoderm reference VCF. Stage 3: FVB/N strain-specific SNPs and indels were removed using a remapped FVB reference VCF. Stage 4: stringent quality thresholds were applied (QUAL ≥ 50, GQ ≥ 30, DP ≥ 30, AD[ALT] ≥ 5). Stage 5: variants were merged within each cohort and compared by bcftools isec to identify cohort-exclusive variants present in all four mice of one genotype and absent from all four of the other. Final output: 4,078 KO-exclusive and 269 WT-exclusive variants.

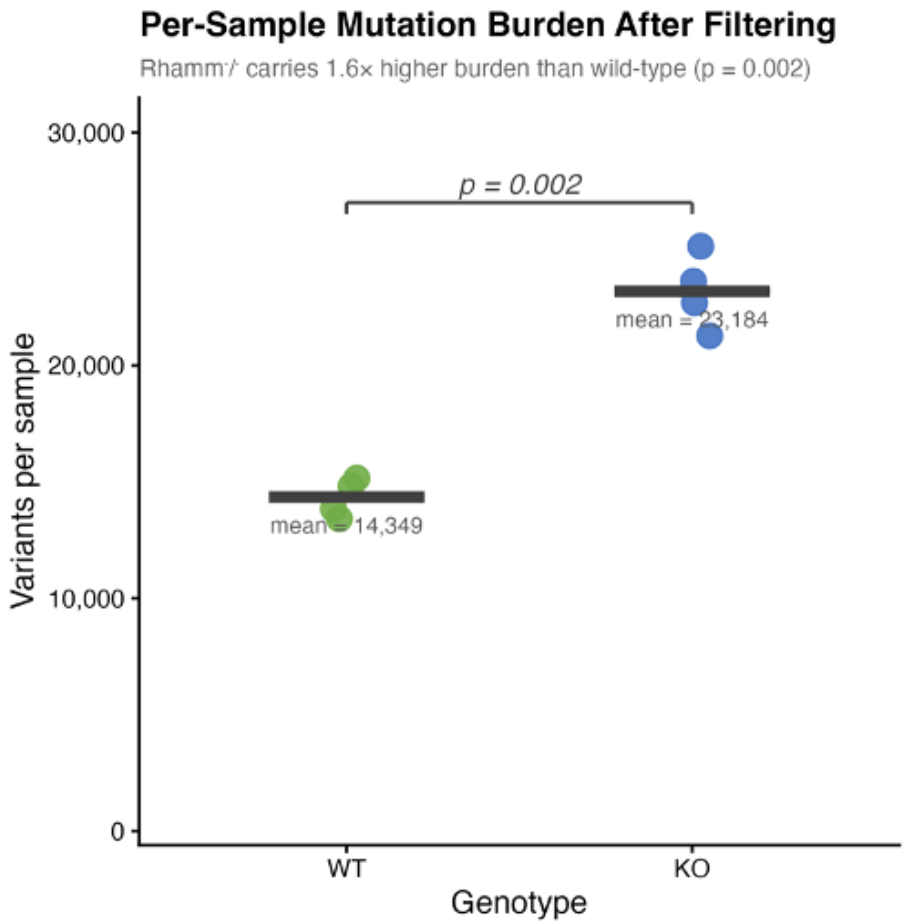

Figure 4. Per-sample variant burden following multi-stage filtering. Strip plot showing the number of high-confidence filtered variants per individual sample, grouped by genotype. Each point represents one mouse (n=4 per group). Horizontal bars indicate the cohort mean. *Rhamm*$^{-/-}$ (KO, blue) samples had a mean of 23,159 variants per sample compared to 14,349 in *Rhamm*$^{+/+}$ (WT, green) samples, representing a 1.6-fold higher mutation burden (two-sample t-test, p = 0.002). Despite this higher overall mutation load, KO samples displayed lower inter-animal variability in per-sample counts. Statistical comparison bracket indicates p-value from a two-sample t-test.

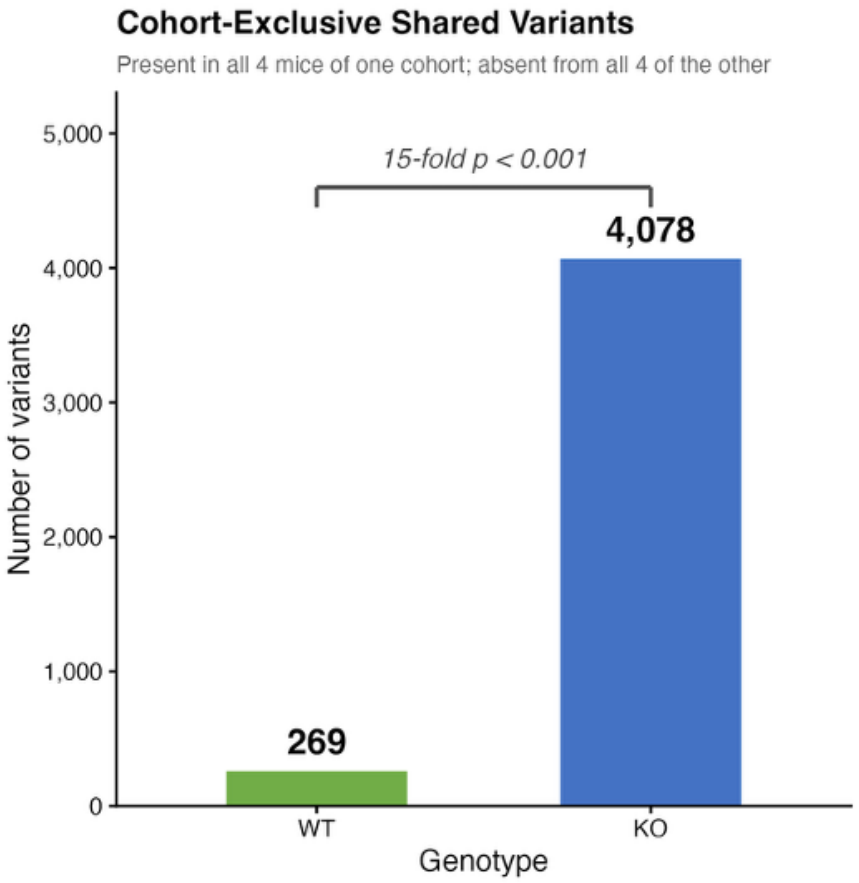

Figure 5. RHAMM-deficient metastases harbour markedly more cohort-shared exclusive variants than wild-type controls. Bar chart showing the number of variants shared across all four mice within each cohort and absent from all four mice of the other cohort. *Rhamm*$^{-/-}$ (KO, blue) mice shared 4,078 cohort-exclusive variants, compared to 269 in *Rhamm*$^{+/+}$ (WT, green) mice; a 15-fold difference. This distribution is highly inconsistent with random chance (binomial test under null hypothesis of equal split, $p < 2.2 \times 10^{-16}$). The disproportionate enrichment of cohort-shared variants in KO animals is consistent with clonal selection toward a dominant genotype across all RHAMM-deficient metastases.

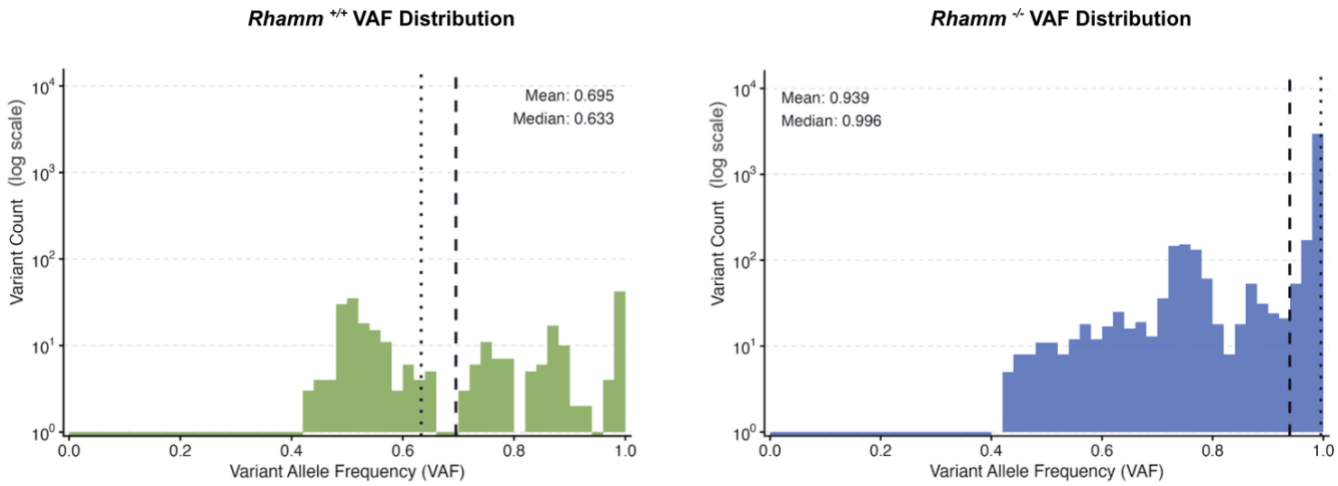

Figure 6. Variant allele frequency distributions of cohort-exclusive variants. Histograms showing the distribution of variant allele frequencies (VAF) for *Rhamm*$^{+/+}$ exclusive variants **(left, green)** and *Rhamm*$^{-/-}$ exclusive variants **(right, blue)** . Y-axes are on a $\log_{10}$ scale to accommodate the wide range of variant counts across VAF bins. Dashed vertical lines indicate mean (black) and median (dotted) VAF. KO-exclusive variants were strongly enriched at high VAF, with a mean of 0.939 and median of 0.996; 82% had VAF ≥ 0.8. WT-exclusive variants showed a broader distribution with mean VAF of 0.695 and median of 0.633; only 34% had VAF ≥ 0.8. VAF distributions between groups were significantly different (Kolmogorov-Smirnov test, $D = 0.61$, $p = 1.22 \times 10^{-79}$), consistent with clonal dominance in KO metastases and ongoing subclonal diversification in WT metastases. VAF values reflect variant frequency across bulk tissue and may underestimate true clonal frequency within the metastatic cell compartment.

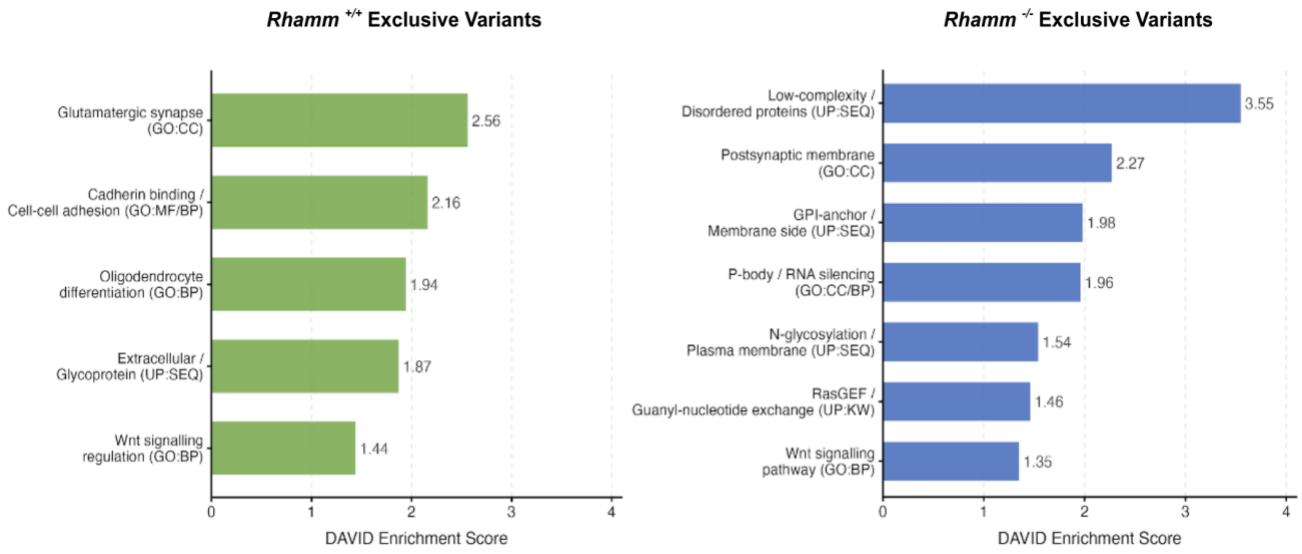

Figure 7. DAVID functional annotation clustering of cohort-exclusive variant gene sets. Horizontal bar charts showing the top DAVID functional annotation clusters for genes harbouring *Rhamm*$^{+/+}$ exclusive variants **(left, green)** and *Rhamm*$^{-/-}$ exclusive variants **(right, blue)**. Bar length represents the cluster enrichment score (geometric mean of $-\log_{10}$ p-values across all terms within a cluster). Representative terms for each cluster are displayed on the y-axis, with annotation database source noted in parentheses (GO:CC, Gene Ontology Cellular Component; GO:MF, Molecular Function; GO:BP, Biological Process; UP:SEQ, UniProt Sequence Features; UP:KW, UniProt Keywords). WT-exclusive genes were most enriched for glutamatergic synaptic terms (score = 2.56; likely attributable to residual FVB × B6 strain background variation) and cadherin-mediated adhesion and morphogenesis (score = 2.16). KO-exclusive genes were most enriched for low-complexity and intrinsically disordered protein regions (score = 3.55), followed by postsynaptic membrane (score = 2.27), GPI-anchored membrane proteins (score = 1.98), RNA silencing and P-body functions (score = 1.96), RasGEF/GTPase signalling (score = 1.46), and Wnt signalling (score = 1.35). Full gene lists and statistical details are provided in Tables 6 and 7. These results are exploratory and require orthogonal validation.

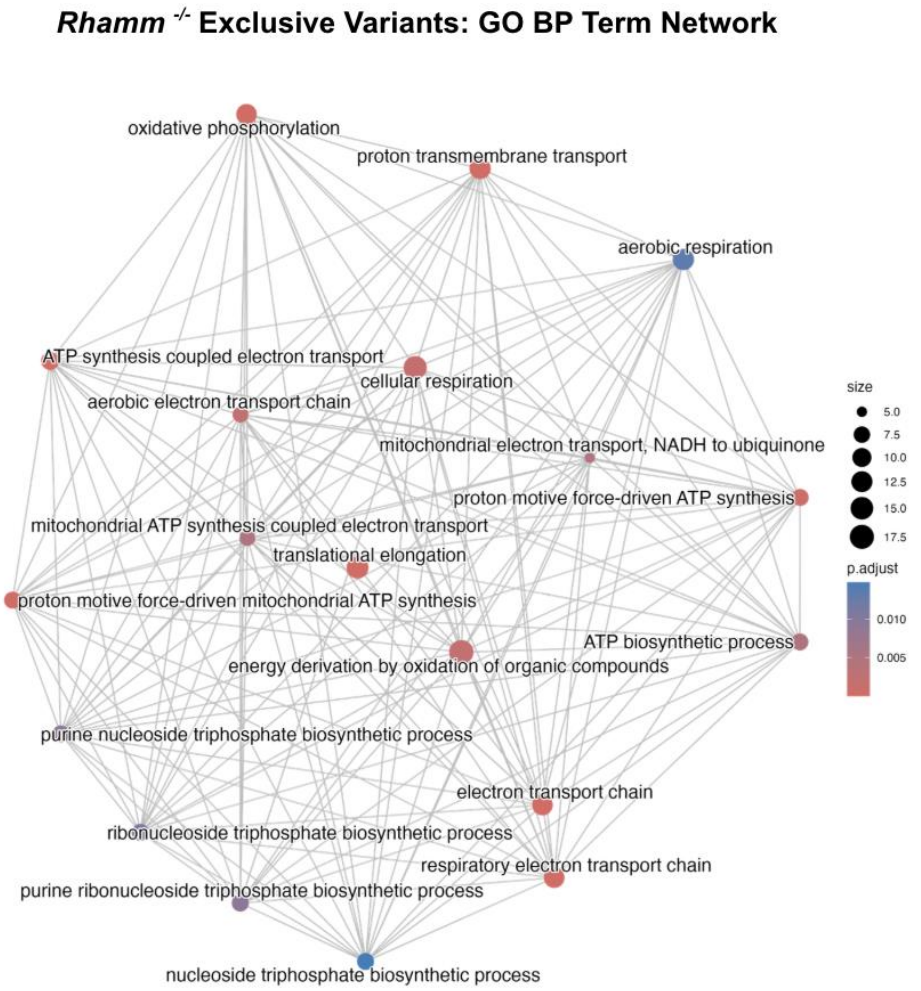

Figure 8. Gene Ontology Biological Process term network for *Rhamm*$^{-/-}$ exclusive variant genes. Network diagram generated by ClusterProfiler (v4.18.4) depicting enriched GO Biological Process terms among genes harbouring *Rhamm*$^{-/-}$ (KO)-exclusive variants. Each node represents a significantly enriched GO term (Benjamini-Hochberg adjusted $p \leq 0.05$); node size is proportional to the number of genes annotated to that term, and node color reflects the adjusted p-value (scale shown at right). Edges connect terms that share gene members, illustrating the relatedness of enriched biological processes. The network is dominated by a highly interconnected cluster of mitochondrial energy metabolism and oxidative phosphorylation terms, including aerobic respiration, ATP synthesis, and the electron transport chain, suggesting that KO-exclusive variants are disproportionately represented in genes associated with cellular energy metabolism and mitochondrial function. These results are concordant with DAVID findings and are presented as a supplementary validation

Tables

Table 1. DNA Extraction and Quality Control Summary.

Genotype	Mouse ID	NanoDrop Conc. (ng/µL)	[a]A260/280	[b]DNA Yield (ng)	[c]TapeStation Conc. (ng/µL)	DNA Used for Lib Prep (µg)	[d]Genome Equiv.	[e]DIN
Rhamm$^{-/-}$	42_1	150	1.85	19,500	15.5	3.0	2.79 × 10⁶	9.9
	7_4	152	1.89	19,760	15.9	3.0	2.82 × 10⁶	9.9
	7_5	170	1.85	22,100	16.1	3.0	3.16 × 10⁶	9.9
	7_6	150	1.85	19,500	16.8	3.0	2.79 × 10⁶	9.9
Rhamm$^{+/+}$	14_8	222	1.87	28,860	14.1	3.0	4.12 × 10⁶	9.9
	27_7	150	1.86	19,500	14.2	3.0	2.79 × 10⁶	9.9
	4_2	150	1.85	19,500	15.4	3.0	2.79 × 10⁶	9.9
	4_4	200	1.86	26,000	15.7	3.0	3.71 × 10⁶	9.9

All eight MMTV-PyMT mice used in this study are grouped by genotype. *Rhamm*$^{-/-}$ (KO, n=4) and *Rhamm*$^{+/+}$ (WT, n=4) cohorts were maintained on a mixed FVB/N × C57BL/6 background. Bulk lung tissue containing metastatic lesions was collected from each animal; tissue input mass ranged from 30–40 mg per sample. All samples passed acceptance thresholds prior to library preparation (A260/280 ≥ 1.8; DIN ≥ 9). [a] NanoDrop A260/280 purity ratio; target ≥ 1.8. Values of 1.85–1.89 across all samples indicate minimal protein contamination. [b] DNA yield calculated as NanoDrop concentration × elution volume (130 µL nuclease-free water per sample). [c] TapeStation concentration represents an independent fluorometric measurement from the submitted QC aliquot, distinct from the NanoDrop estimate. [d] Genome equivalents estimated as DNA mass input (3.0 µg) ÷ diploid mouse genome mass (6.4 pg per cell). [e] DNA Integrity Number (DIN; scale 1–10). All eight samples scored 9.9, the highest value consistently reported by the TapeStation system, confirming highly intact, non-degraded genomic DNA suitable for PacBio HiFi library preparation. Libraries were prepared using 3.0 µg per sample with the SMRTbell Express Template Prep Kit 2.0.

Table 2. Sequencing Quality Metrics.

Genotype	Mouse ID	Total Reads	Mean Read Length (bp)	N50 (bp)	Mean Coverage	Mean Read Quality (Q)
Rhamm$^{-/-}$	42_1	9,857,604	17,511	18,082	48.2×	23.4
	7_4	9,232,742	19,820	20,503	50.4×	23.6
	7_5	9,670,742	18,977	19,562	51.5×	23.2
	7_6	8,858,419	21,123	22,063	51.5×	23.2
Rhamm$^{+/+}$	14_8	9,666,715	17,559	17,915	48.8×	24.1
	27_7	9,005,361	19,522	20,206	47.8×	22.5
	4_2	8,867,877	19,508	20,171	46.7×	22.9
	4_4	8,607,484	17,688	18,141	42.2×	22.8

Library preparation and per-sample PacBio HiFi sequencing statistics following CCS generation and alignment to GRCm39/mm39. All samples used 3.0 µg HMW genomic DNA as input for the SMRTbell Express Template Prep Kit 2.0 on the PacBio Sequel IIe platform (target coverage: 25–40×). Total reads, mean read length, and N50 were obtained from NanoPlot; mean depth and breadth of coverage were calculated using mosdepth. Q23 corresponds to ~99.5% base accuracy. All samples exceeded the 25× minimum target coverage for SNV calling; coverage at ≥30× ranged from 85% (4_4) to 91% (7_4, 7_5, 7_6, 14_8). Grey shading = *Rhamm*$^{-/-}$; White Shading = *Rhamm*$^{+/+}$. NanoPlot interactive HTML reports are available in the GitHub repository at https://github.com/audreygoddard04/RHAMM-HiFi-lung-metastasis.

Table 3. mosdepth Coverage Summary

Genotype	Mouse ID	≥1× (%)	≥5× (%)	≥10× (%)	≥30× (%)	≥50× (%)	Median Depth (×)	Mean Depth (×)
Rhamm$^{-/-}$	42_1	94	94	94	90	47	49	48.19
	7_4	94	94	94	91	57	52	50.36
	7_5	94	94	94	91	61	53	51.53
	7_6	94	94	94	91	58	52	51.48
Rhamm$^{+/+}$	14_8	94	94	94	91	51	50	48.77
	27_7	94	94	94	90	46	49	47.83
	4_2	94	94	94	89	40	47	46.67
	4_4	94	94	94	85	24	43	42.20

Breadth of coverage (percentage of genome covered at each depth threshold) and depth statistics per sample, calculated using mosdepth on sorted, indexed BAM files aligned to GRCm39/mm39. Breadth at ≥1×, ≥5×, and ≥10× was uniformly 94% across all samples, reflecting the proportion of the reference genome accessible with these read depths. Coverage at ≥50× was more variable (24%–61%), consistent with the uneven read depth distribution typical of PacBio long-read data. Median depth reflects the central tendency of per-base coverage and is less sensitive to extreme outliers than mean depth. Grey shading = *Rhamm*$^{-/-}$; White shading = *Rhamm*$^{+/+}$.

Table 4. WT-Exclusive Genes DAVID Functional Annotation Clustering

Cluster	Enrichment Score	Category	Representative Term	Gene Count	Fold Enrichment	p-value	Sig.	Representative Genes
1	2.56	GO:CC	Glutamatergic synapse	13	6.84	2.47E-07	***	*ROBO2, NTRK2, CNTNAP2, SLC6A17, LRRK2, CTNND2, NRG1, BTBD9, CDH9, RIMS2, CACNB4, PTPRZ1, NBEA*
		GO:CC	Dendrite	9	6.53	5.68E-05	***	*NTRK2, CNTNAP2, PTPRZ1, MME, BGLAP, CTNND2, LRRK2, NRG1, CDH9*
		GO:CC	Axon	8	6.96	1.27E-04	***	*NTRK2, CNTNAP2, PTPRZ1, MME, DCC, LRRK2, NRG1, CDH9*
		GO:CC	Neuronal cell body	7	5.06	2.30E-03	**	*NTRK2, CNTNAP2, PTPRZ1, MME, LRRK2, SLC1A3, NRG1*
		GO:CC	Neuron projection	5	6.18	8.18E-03	**	*RIMS2, CNTNAP2, LRRK2, SLC1A3, NRG1*
		GO:CC	Postsynapse	5	5.37	1.32E-02	*	*NTRK2, CNTNAP2, SLC6A17, LRRK2, NRG1*
2	2.16	GO:BP	Cell morphogenesis	5	16.36	2.34E-04	***	*CNTNAP2, CDH12, NRG1, CDH9, CDH18*
		GO:MF	Cadherin binding	4	25.31	5.09E-04	***	*CTNND2, CDH12, CDH9, CDH18*
		GO:BP	Cell-cell adhesion	5	10.78	1.12E-03	**	*DCC, CTNND2, CDH12, CDH9, CDH18*
		GO:BP	Cell-cell junction assembly	3	47.44	1.76E-03	**	*CDH12, CDH9, CDH18*
		GO:BP	Adherens junction organization	3	41.86	2.26E-03	**	*CDH12, CDH9, CDH18*
		GO:CC	Catenin complex	3	45.21	1.94E-03	**	*CDH12, CDH9, CDH18*
		INTERPRO	Cadherin domain	3	42.10	2.22E-03	**	*CDH12, CDH9, CDH18*
		GO:MF	Beta-catenin binding	4	14.71	2.44E-03	**	*CTNND2, CDH12, CDH9, CDH18*

Cluster	Enrichment Score	Category	Representative Term	Gene Count	Fold Enrichment	p-value	Sig.	Representative Genes
		GO:CC	Adherens junction	4	10.32	6.58E-03	**	*CTNND2, CDH12, CDH9, CDH18*
3	1.94	GO:BP	Oligodendrocyte differentiation	3	24.12	6.68E-03	**	*NTRK2, PTPRZ1, NRG1*
		GO:BP	Positive regulation of neuron projection development	4	10.26	6.68E-03	**	*NTRK2, PTPRZ1, DCC, NRG1*
4	1.87	UP:SEQ	Extracellular topological domain	17	2.94	8.36E-05	***	*ROBO2, NTRK2, CNTNAP2, MME, SLC6A17, DCC, FZD6, SLC1A3, NRG1, ANO6, TRHDE, CDH9, ANO3, PTPRZ1, TMEM117, CDH12, SERINC5*
		UP:KW	Glycoprotein	24	1.75	1.60E-03	**	*ROBO2, NTRK2, CNTNAP2, MME, SLC6A17, CPQ, DCC, TUSC3, COL22A1, CTNND2, FZD6, SLC1A3, NRG1, ANO6, TRHDE, CDH9, ANO3, MGAT4C, PTPRZ1, CDH12, CDH18 (+3 more)*
		GO:CC	Plasma membrane	22	1.63	1.35E-02	*	*ROBO2, NTRK2, MME, SLC6A17, DCC, LRRK2, CTNND2, FZD6, SLC1A3, NRG1, ANO6, CDH9, ANO3, CACNB4, PTPRZ1, NBEA, TMEM117, CDH12, CDH18 (+3 more)*
5	1.44	GO:BP	Regulation of canonical Wnt signalling pathway	3	44.48	2.01E-03	**	*DAAM2, CTNND2, LRRK2*
		INTERPRO	ARM-type fold	5	5.88	9.63E-03	**	*STAG1, NBEA, DAAM2, CTNND2, LRRK2*

Gene symbols (Mus musculus) corresponding to the 269 WT-exclusive variants were submitted to DAVID. Enrichment score = geometric mean of $-\log_{10}$(p-values) across all terms in a cluster; higher scores reflect stronger overrepresentation. Fold enrichment = (observed/expected) ratio for each term. Category abbreviations: GO:BP = Gene Ontology Biological Process; GO:CC = Cellular Component; GO:MF = Molecular Function; INTERPRO = InterPro protein domain database; UP:KW = UniProt Keywords; UP:SEQ = UniProt Sequence Features; KEGG = Kyoto Encyclopedia of Genes and Genomes. Only the top five clusters with enrichment score > 1.0 are shown. p-values are from Fisher's exact test (EASE score) implemented in DAVID. Significance: *** $p \leq 0.001$; ** $p \leq 0.01$; * $p \leq 0.05$; — not significant.

Table 5. KO-Exclusive Genes DAVID Functional Annotation Clustering

Cluster	Enrichment Score	Category	Representative Term	Gene Count	Fold Enrichment	p-value	Sig.	Representative Genes
1	3.55	UP:SEQ	Low-complexity / compositionally biased regions (polar residues)	90	1.45	6.24E-05	***	*ALK, VEZT, CRYBG2, TENM2, SH3KBP1, CHD9, NRK, SCHIP1, PCDH11X, GOLIM4, AFF3, ARHGAP6, SYNE1, GRM5, PCLO, PSD3, FBRSL1, NCKAP5, TNS3, SYBU, RBFOX1, KLF12, EPHA6 (+67 more)*
		UP:SEQ	Intrinsically disordered / low-complexity regions (basic/acidic)	88	1.29	4.23E-03	**	*ALK, VEZT, CRYBG2, TENM2, EDA, SH3KBP1, CHD9, NRK, SCHIP1, AFF3, SYNE1, GRM5, PCLO, PSD3, FBRSL1, NCKAP5, TNS3, SYBU, RBFOX1, LMO7, RAPGEF2, RAPGEF6, COL11A1 (+65 more)*
		UP:SEQ	REGION: Disordered	160	1.11	2.47E-02	*	*VEZT, NRK, RTKN2, PCDH11X, GOLIM4, YARS1, DCAF5, FBRSL1, NCKAP5, TNS3, SYBU, RBFOX1, EPHA6, LMO7, AFF3, RAPGEF2, RAPGEF6, OSBP2, KHDRBS1, MYCBP2, ITPR2, PCDH17, PRDM16 (+137 more)*
2	2.27	GO:CC	Postsynaptic membrane	12	3.82	3.41E-04	***	*GRIA1, GRM5, TENM2, GLRA2, HTR7, NBEA, LRRTM4, SLC1A7, GABRG2, PCDH17, SYNE1, KCNH1*
		GO:CC	Synapse	21	1.93	6.46E-03	**	*GRIA1, TENM2, CNTN5, SH3KBP1, SLC1A7, GABRG2, EXT1, GLRA2, DAB1, HTR7, PCLO, NBEA, LRRTM4, RPS6KA1, PSD3, RAPGEF2, VPS54, GPC6, WASF2, CACNG4, KCNH1*

Cluster	Enrichment Score	Category	Representative Term	Gene Count	Fold Enrichment	p-value	Sig.	Representative Genes
3	1.98	UP:SEQ	GPI-anchor amidated serine	6	8.15	8.39E-04	***	*NTNG1, RAET1D, RAET1E, CNTN5, GPC6, RECK*
		GO:CC	Side of membrane	7	4.51	4.73E-03	**	*NTNG1, RAET1D, RAET1E, CNTN5, MDGA2, GPC6, RECK*
		UP:KW	GPI-anchor	7	3.44	1.62E-02	*	*NTNG1, RAET1D, RAET1E, CNTN5, MDGA2, GPC6, RECK*
4	1.96	GO:CC	P-body	7	7.08	4.75E-04	***	*ZC3H12D, LIN28A, AGO3, AGO4, EIF4E, SYNE1, PATL2*
		GO:BP	Pre-miRNA processing	3	21.24	8.46E-03	**	*LIN28A, AGO3, AGO4*
		UP:KW	RNA-mediated gene silencing	4	6.83	2.03E-02	*	*SPOCD1, LIN28A, AGO3, AGO4*
		GO:BP	Regulatory ncRNA-mediated gene silencing	4	6.29	2.56E-02	*	*SPOCD1, LIN28A, AGO3, AGO4*
5	1.54	UP:SEQ	N-linked glycosylation (GlcNAc… Asn)	65	1.58	1.16E-04	***	*ALK, CD164L2, TENM2, WNT2B, EDA, CSF3R, GOLIM4, PTPRK, ADRA1B, TSPAN11, GRM5, HTR7, SGCD, LRRTM4, RSPO2, HAVCR1, KCNH1, CSGALNACT1, H60B, EPHA6, GABRG2, SLC9A7, CDH12, ROR2, GRIA1 (+40 more)*
		GO:CC	Plasma membrane	91	1.42	1.85E-04	***	*RAB3B, ALK, VEZT, TENM2, EDA, CSF3R, SCHIP1, RTKN2, PCDH11X, PTPRK, ADRA1B, TSPAN11, GRM5, HTR7, SGCD, KCNT2, LRRTM4, PSD3, HAVCR1, KCNH1, EPHA6, GABRG2, SLC9A7, CDH12, ROR2, GRIA1 (+65 more)*
		UP:KW	Glycoprotein	73	1.23	2.43E-02	*	*ALK, CD164L2, TENM2, WNT2B, EDA, CSF3R, PCDH11X,*

Cluster	Enrichment Score	Category	Representative Term	Gene Count	Fold Enrichment	p-value	Sig.	Representative Genes
								GOLIM4, PTPRK, ADRA1B, TSPAN11, GRM5, HTR7, SGCD, LRRTM4, RSPO2, HAVCR1, KCNH1, CSGALNACT1, H60B, EPHA6, GABRG2, CDH12, ROR2, GRIA1 (+48 more)
9	1.46	UP:KW	Guanine-nucleotide releasing factor (RasGEF)	7	3.92	8.78E-03	**	*BCAR3, PLEKHG1, PSD3, MYCBP2, RAPGEF2, DOCK2, RAPGEF6*
		GO:MF	Guanyl-nucleotide exchange factor activity	7	2.87	3.59E-02	*	*BCAR3, PLEKHG1, PSD3, MYCBP2, RAPGEF2, DOCK2, RAPGEF6*
		UP:SEQ	DOMAIN: Ras-GEF	3	9.75	3.74E-02	*	*BCAR3, RAPGEF2, RAPGEF6*
10	1.35	GO:BP	Positive regulation of canonical Wnt signalling pathway	6	5.22	5.96E-03	**	*EDA, CAPRIN2, RSPO2, ROR2, RECK, RNF220*
		GO:BP	Canonical Wnt signalling pathway	5	4.39	2.74E-02	*	*EXT1, EDA, WNT2B, RSPO2, RECK*
		KEGG	Wnt signalling pathway	5	2.97	8.54E-02	—	*WNT2B, RSPO2, PRICKLE2, ROR2, SEMP2L2B*

Gene symbols (Mus musculus) corresponding to the 4,078 KO-exclusive variants were submitted to DAVID. Clusters shown: 1 (low-complexity/disordered proteins, score 3.55), 2 (postsynaptic membrane, score 2.27), 3 (GPI-anchored membrane proteins, score 1.98), 4 (RNA silencing/P-bodies, score 1.96), 5 (membrane glycoproteins/N-glycosylation, score 1.54), 9 (RasGEF/GTPase signalling, score 1.46), and 10 (Wnt signalling, score 1.35). All other notation as in Table 6. The higher enrichment score for Cluster 1 (3.55 vs. WT Cluster 1 score 2.56) and the distinct functional programs (RNA silencing, RasGEF/Wnt signalling, membrane glycoproteins) relative to WT-exclusive clusters (cadherin adhesion, neuronal morphogenesis) suggest divergent mutational trajectories between RHAMM-expressing and RHAMM-deficient metastases.

Table 6. Cohort Variant Comparison Summary

Metric	KO (*Rhamm*$^{-/-}$)	WT (*Rhamm*$^{+/+}$)	Source / Tool
Cohort-exclusive variants (final)	4,078	269	bcftools isec
Homozygous ALT (GT=1/1) among exclusive	81%	21%	bcftools; genotype filter
% exclusive variants with VAF ≥ 0.8	82%	34%	VAF distribution
Mean VAF of exclusive variants	0.939	0.695	AD[ALT]/DP calculation
Mean variants per sample (pre-merge)	23,159	14,349	Per-sample VCF count
Transition:Transversion ratio (Ti/Tv)	1.34	1.46	bcftools stats
Variant Type	3,642 SNV / 263 INS / 173 DEL	239 SNV / 20 INS / 10 DEL	DeepVariant

Summary of key variant metrics comparing *Rhamm*$^{-/-}$ (KO) and *Rhamm*$^{+/+}$ (WT) cohort-exclusive variant sets.

Table 7. Raw pre-filter variant calls from DeepVariant across all eight samples.

Genotype	Mouse ID	Total Variants	SNVs	Indels	Ts/Tv
Rhamm$^{-/-}$	42_1	9,857,604	17,511	18,082	48.2×
	7_4	9,232,742	19,820	20,503	50.4×
	7_5	9,670,742	18,977	19,562	51.5×
	7_6	8,858,419	21,123	22,063	51.5×
Rhamm$^{+/+}$	14_8	9,666,715	17,559	17,915	48.8×
	27_7	9,005,361	19,522	20,206	47.8×
	4_2	8,867,877	19,508	20,171	46.7×
	4_4	8,607,484	17,688	18,141	42.2×

Raw unfiltered variant counts prior to any filtering stages. The majority of these variants represent germline polymorphisms, strain-background SNPs, and low-confidence calls subsequently removed by the multi-stage filtering pipeline (Figure 3). SNV and indel counts may not sum exactly to total variants due to multiallelic site representation.

Supplementary Material

Appendix A. Glossary

Term	Definition
Alignment	The process of mapping sequencing reads to a reference genome to determine their genomic origin.
Average Read Length	The mean length, in base pairs, of sequencing reads generated by a sequencing platform.
BAM (Binary Alignment Map) File	A binary, indexed file format that stores aligned sequencing reads.
Bioinformatics Pipeline	An ordered series of computational steps that process raw sequencing data into interpretable results.
Bulk Sequencing	Sequencing performed on DNA extracted from a mixed population of cells, capturing average and heterogeneous signals.
CCS (Circular Consensus Sequencing)	A process that generates highly accurate long reads (HiFi reads) by combining multiple passes of the same DNA molecule within a SMRTbell library.
Centrosome	A cellular structure involved in mitotic spindle formation; instability can contribute to chromosomal missegregation.
Chromosomal Instability (CIN)	A high rate of chromosome missegregation leading to gains, losses, or rearrangements of chromosomes.
CLI (Command-Line Interface)	A text-based interface used to run software and automate workflows.
Clonal Dominance	The expansion of one or a few genetically similar tumor cell populations within a heterogeneous tumor.
Clonal Heterogeneity	The presence of multiple genetically distinct cell populations within a tumor.
COSMIC Mutational Signatures	Characteristic patterns of mutations associated with specific DNA damage or repair processes, curated in the COSMIC database.
Coverage (Genome Coverage)	The proportion of the genome represented by sequencing reads and the depth at which it is sequenced.
De Novo Assembly	Reconstruction of a genome or genomic region without reliance on a reference sequence.

DIN (DNA Integrity Number)	A numerical assessment of DNA quality ranging from 1 to 10, where higher numbers indicate less degraded DNA.
Epigenetic Heterogeneity	Variation in epigenetic marks, such as DNA methylation, across tumor cells.
ERK Signaling Pathway	A mitogen-activated protein kinase pathway involved in cell proliferation, survival, and motility.
FastQC	A tool for quality control analysis of sequencing reads, identifying issues such as low-quality bases or adapter contamination.
FASTQ File	A text-based file format containing raw sequencing reads and their quality scores.
Fisher's Exact	A statistical test used to determine if there is a non-random association between two categorical variables (e.g., mutation presence vs. genotype).
Genome-Wide	Analysis encompassing the entire genome rather than selected regions.
HAS Enzymes (Hyaluronan Synthases)	Enzymes responsible for the synthesis of hyaluronan.
HiFi Reads (PacBio HiFi)	Highly accurate long sequencing reads generated by PacBio circular consensus sequencing.
HiFiCNV	A software tool designed to identify copy number variations (CNVs) from PacBio HiFi sequencing data.
High-Molecular-Weight (HMW) DNA	Long, intact DNA fragments typically >10 kb required for long-read sequencing.
HiPhase	A tool for jointly phasing small variants (SNVs), structural variants (SVs), and tandem repeats specifically for PacBio data into haplotypes.
Hyaluronan (HA)	A glycosaminoglycan component of the extracellular matrix involved in cell signaling and motility.
Intrametastatic Heterogeneity	Genetic or epigenetic diversity within a single metastatic lesion.
isec (trio)	A tool (part of bcftools/samtools) used to find the intersection or difference between multiple VCF files, often used in trio analysis to identify de novo mutations or inheritance patterns.

J-statistic (Youden's J)	A statistical measure for comparing mutation patterns or clusters and determining the optimal threshold for a test by balancing sensitivity and specificity.
Linux	A free, open-source, Unix-like operating system kernel that manages hardware resources such as processes, memory, and storage.
Long-Read Sequencing	Sequencing technologies that generate reads thousands to tens of thousands of bases long.
Lung Metastases	Secondary tumor lesions formed when cancer cells colonize lung tissue.
MMTV-PyMT Mouse Model	A transgenic mouse model expressing the Polyoma Middle T antigen under the MMTV promoter, used to study breast cancer metastasis.
mosdepth	A high-speed tool for fast calculation of genome-wide coverage from BAM files.
MultiQC	A reporting tool that aggregates results from multiple bioinformatics tools (FastQC, samtools stats, mosdepth, etc.) into a single, interactive HTML report.
MutSeqR	An R package used for standardized analysis of error-corrected sequencing data, including variant annotation, filtering, and prioritization in mutation analysis.
Mutation Burden	The total number of mutations present in a genome or genomic region.
Nanobind DNA Extraction	A method for isolating ultra–high-molecular-weight DNA using nanostructured silica disks.
NanoComp	A tool used to compare quality and read length distributions across multiple long-read sequencing samples.
NanoPlot	A visualization tool designed for long-read data to plot read lengths and quality scores.
PacBio (Pacific Biosciences)	A sequencing platform specializing in long-read technologies.
pbmm2	A PacBio-optimized wrapper for the minimap2 aligner, used to map HiFi reads to a reference genome and generate sorted BAMs.
pbsv (discover & call)	A PacBio suite for identifying potential structural variant (SV) signatures and genotyping them.
plotCoverage	A tool (part of deepTools) used to assess sequencing depth by plotting a histogram of coverage frequency across the genome.

Purity Ratio (260/280)	A spectrophotometric measure of DNA purity indicating protein contamination levels.
Rainfall Plot	A visualization showing the genomic distribution of mutations across chromosomes.
Read Depth	The number of sequencing reads covering a given genomic position.
Reference Genome	The standard mouse genome assembly used for sequence alignment.
RHAMM	Receptor for Hyaluronan-Mediated Motility; a protein involved in cell motility, mitosis, and signaling, implicated in cancer progression.
Rhamm$^{+/+}$	Wild-type genotype with intact RHAMM expression.
Rhamm$^{-/-}$	RHAMM-deficient genotype, lacking expression.
samtools (sort/index/stats)	A toolkit for manipulating BAM/SAM files, including sorting, indexing, and generating comprehensive statistics.
sawfish (discover & call)	A specialized caller for identifying structural variations and somatic mutations in long-read datasets.
Scoping Review	A systematic approach to mapping existing literature and identifying knowledge gaps.
Sequencing Coverage Depth	Average number of times each base in the genome is sequenced.
Short Read Eliminator (SRE)	A size-selection step that removes short DNA fragments prior to long-read sequencing.
SigProfiler (Extractor/Assignment/Clusters)	A suite for extracting de novo mutational signatures, assigning them to known COSMIC signatures, and detecting clustered mutations (kataegis).
Single-Nucleotide Variant (SNV)	A substitution of a single nucleotide at a specific genomic position.
SSH (Secure Shell)	A protocol for secure remote access to computational servers.
Structural Variant (SV)	Large-scale genomic alterations including insertions, deletions, inversions, or translocations.
SVHawkeye	An ultra-fast visualization software used to inspect structural variant calls from long-read data.
TapeStation	An automated electrophoresis platform used to analyze the size, concentration, and integrity of DNA and RNA samples.

TRGT (Tandem Repeat Genotyping Tool)	A tool to characterize and genotype tandem repeat loci and repeat expansions from HiFi reads.
Tumor Microenvironment	The surrounding stromal cells, immune cells, and extracellular matrix influencing tumor behavior.
Ultra–High-Molecular-Weight (uHMW) DNA	Extremely long DNA fragments requiring careful handling to prevent shearing.
Variant Call Format (VCF)	A standardized file format for storing genetic variant information.
Variant Calling	The process of identifying genetic variants relative to a reference genome.

Appendix B. Search Strategy Design

Concept 1: Genetic Architecture, Environment, and Entities	
MeSH Term/ Subject Headings	Genetic Heterogeneity/ or DNA Copy Number Variations/ or exp Genomic Instability/ or Tumor Microenvironment or Hyaluronan Receptors/ or Hyaluronic Acid/ or Hyaluronan Synthases/ or Hyaluronoglucosaminidase/ or exp Hyalectins/
Keywords	Genetic Heterogeneit*.mp. or Tumor Heterogeneit*.mp. or Clonal Heterogeneit*.mp. or Clonal Diversit*.mp. or Somatic Varia*.mp. or Mosaicism*.mp. or Allelic Heterogeneit*.mp. or Chimerism*.mp. or epigenetic Heterogeneit*.mp. or Copy Number Variation*.mp. or CNV.ti,ab. or Genetic Diversit*.mp. or Locus Heterogeneit*.mp. or Phenotypic Heterogeneit*.mp. or Intertumor Heterogeneit*.mp. or Intratumor Heterogeneit*.mp. or Population Heterogeneit*.mp. or Cellular Heterogeneit*.mp. or Positional Heterogeneit*.mp. or Temporal Heterogeneit*.mp. or Subclonal Diversit*.mp. or Microheterogeneit*.mp. or Microenvironment*.mp. or Hyaluron*.mp. or RHAMM.ti,ab. or IHABP.ti,ab. or CD168.ti,ab. or CD44.ti,ab. or Chondroitin Sulfate Proteoglycan*.mp. or hasA Enzyme*.mp. or HAS1.ti,ab. or HAS2.ti,ab. or HAS3.ti,ab. or "Hyals".ti,ab. or Amvisc*.mp. or Luronit*.mp. or Hyvisc*.mp. or Healon*.mp. or Etamucine*.mp. or Amo Vitrax*.mp. or Biolon*.mp. or Aggrecan*.mp. or Brevican*.mp. or Neurocan*.mp. or Versican*.mp.

Concept 2: Breast Neoplasm	
MeSH Term/ Subject Headings	exp Breast Neoplasms/ or Paget's Disease, Mammary/ or exp Mammary Neoplasms, Animal/
Keywords	Breast Neoplasm*.mp. or Breast Tumor*.mp. or Breast Cancer*.mp. or Breast Malignant Neoplasm*.mp. or Breast Malignant Tumor*.mp. or Mammary Cancer*.mp. or Mammary Neoplasm*.mp. or Breast Carcinoma*.mp. or Mammary Carcinoma*.mp. or Mammary Tumor*.mp. or Lobular carcinoma*.mp. or "Hereditary Breast and Ovarian Cancer Syndrome".mp. or HBOC.ti,ab. or Ductal Carcinoma*.mp. or Paget's Disease*.mp. or TNBC*.ti,ab. or Mucinous Adenocarcinoma*.mp. or Mucinous Carcinoma*.mp. or Papillary carcinomas of the breast*.mp. or PCB.ti,ab. or tubular carcinoma*.mp. or Cribriform Carcinoma*.mp. or Medullary Carcinoma*.mp. or Colloid Carcinoma*.mp.

Concept 3: Prognosis Factors	
MeSH Term/ Subject Headings	exp Prognosis/ or exp Survival Analysis/ or exp Death/ or Survival/ or exp Vital Statistics/ or exp Quality of Life/ or Cancer Survivors/ or Long-Term Care/ or exp Aftercare/ or exp Carcinogenesis/ or exp

	Neoplasm Metastasis/ or Population Health Management/ or exp Disease Management/ or exp Managed Care Programs/ or exp Remission, Spontaneous/ or exp Recurrence/ or exp Immunotherapy/ or exp Chemoradiotherapy/ or Palliative Care/ or Palliative Medicine/ or Genomic Medicine/ or exp Precision Medicine/ or exp Radiotherapy/ or Neoplasm Recurrence, Local/ or Tumor Burden/ or exp Risk Assessment/ or exp Gene Expression Profiling/
Keywords	Prognos*.mp. or Death*.mp. or End Of Life*.mp. or Gene Expression Profiling*.mp. or Gene Expression Monitor*.mp. or Surviv*.mp. or Life Expectanc*.mp. or Gene Expression Pattern Analys*.mp. or Morbidit*.mp. or Mortalit*.mp. or Fatal Outcome*.mp. or Transcript Expression Analys*.mp. or Transcriptome Analys*.mp. or Quality of Life*.mp. or Aftercare*.mp. or Rehab*.mp. or Tumorigenes*.mp. or Disease Management*.mp. or Oncogenes*.mp. or Carcinogenes*.mp. or Metastas*.mp. or Wellbeing*.mp. or Transcriptome Profiling*.mp. or Gene Expression Analys*.mp. or Managed Care*.mp. or mRNA Differential Display*.mp. or Risk Assessment*.mp. or or Recurrence*.mp. or Relaps*.mp. or Recrudescence*.mp. or Immunotherap*.mp. or Radioimmunotherap*.mp. or Lymphocyte Depletion*.mp. or Transplantation Conditioning*.mp. or Risk Analys*.mp. or Ribosome Profiling*.mp. or RNA-Seq*.mp. or Spontaneous Healing*.mp. or Single-Cell Gene Expression Analysis*.mp. or Subtractive Hybridization Techniques*.mp. or Tumor Burden*.mp. or Tumor Weight*.mp. or Tumor Volume*.mp. or Tumor Load*.mp. or Palliative*.mp. or Genomic Medicine*.mp. or Neoplasm Recurrence*.mp. or Radiotherap*.mp. or Radiation*.mp. or Chemoradiotherap*.mp. or Precision Medicine*.mp. or "P-Health".mp. or Individualized Medicine*.mp. or Personalized Medicine*.mp. or Theranostic*.mp. or Predictive Medicine*.mp.

Concept 4: Clonal Homogeneity	
MeSH Term/ Subject Headings	exp Clone Cells/ or exp Clonal Evolution/ or Clonal Selection, Antigen-Mediated/
Keywords	Clonal Expansion*.mp. or Colony Clonalit*.mp. or Homogeneit*.mp. or Clonal Dynamic*.mp. or Clonal dominan*.mp.

Appendix C. Bioinformatics Pipeline

Key Commands & Spftware Versions

Alignment:

```
pbmm2 align --sort --preset HiFi reference.mmi
sample.hifi.bam sample.aligned.bam
```

BAM indexing and stats:

```
samtools index sample.aligned.bamsamtools flagstat
sample.aligned.bamsamtools stats sample.aligned.bam | grep
'^SN'
```

Variant calling (DeepVariant — GPU):

```
run_deepvariant --model_type PACBIO --ref mm39.fa --reads
sample.aligned.bam \  --output_vcf
sample.deepvariant.vcf.gz --num_shards 48
```

Phasing (HiPhase):

```
hiphase --bam sample.aligned.bam --vcf
sample.deepvariant.vcf.gz \  --output-vcf
sample.phased.vcf.gz --reference mm39.fa
```

Coverage (mosdepth):

```
mosdepth --quantize 0:1:5:10:30:50: --no-abbrev sample
sample.aligned.bam
```

VAF filtering (Stage 1):

```
bcftools filter -i 'FORMAT/VAF[0] >= 0.3'
sample.phased.vcf.gz -Oz -o lowVAFout.sample.vcf.gz
```

Strain filtering (Stage 2):

```
bcftools isec -C sample.vcf.gz B6_12L1_reference.vcf.gz -p
./isec_B6 -Oz
```

Strain filtering (Stage 3):

```
bcftools isec -C sample.vcf.gz FVB_remapped_snps.vcf.gz -p
./isec_FVB -Oz
```

Quality filtering (Stage 4):

```
bcftools filter -i 'QUAL>=50 && FORMAT/GQ[0]>=30 &&
FORMAT/DP[0]>=30 && FORMAT/AD[0:1]>=5' \  input.vcf.gz -Oz
-o quality_filtered.vcf.gz
```

Cohort merge and isec (Stage 5):

```
bcftools merge sample1.vcf.gz sample2.vcf.gz sample3.vcf.gz
sample4.vcf.gz -Oz -o cohort.merged.vcf.gzbcftools isec
KO.merged.vcf.gz WT.merged.vcf.gz -p ./isec_cohorts -Oz
```

Variant counting:

```
bcftools view -H file.vcf.gz | wc -l
```

VAF extraction and mean calculation:

```
bcftools query -f '[%AD]\n' KO-exclusive.vcf.gz \  | awk -F',' '{if($1+$2>0) print $2/($1+$2)}' \  | awk '{sum+=$1; n++} END {print "Mean VAF:", sum/n}'
```

Ti/Tv ratio:

```
bcftools stats KO-exclusive.vcf.gz | grep 'Ts/Tv'
```

Software	Version	Description
FastQC	0.12.1	Raw read quality control; per-base quality scores, GC content, adapter contamination
pbmm2	v1.17	PacBio-optimized minimap2 wrapper for HiFi alignment
samtools	v1.19.2	BAM sorting, indexing, flagstat, stats
DeepVariant	v1.6.0	Deep-learning SNV/indel caller (GPU-accelerated)
HiPhase	v1.6.0	PacBio haplotype phasing tool
bcftools	v1.19	VCF manipulation, isec, merge, filter, stats, query
mosdepth	v0.3.10	Genome-wide coverage calculation
NanoPlot	v1.46.2	Long-read QC visualization
NanoComp	v1.25.6	Multi-sample long-read QC comparison
MultiQC	v1.33	Aggregated QC reporting
SnpEff	v5.4c	Variant functional annotation; predicts effects on genes and proteins
VEP	v115.2	Ensembl Variant Effect Predictor; variant consequence and impact annotation
DAVID	2024 update	Functional annotation clustering
ClusterProfiler	v4.18.4	R package for GO and KEGG pathway enrichment analysis with Benjamini-Hochberg correction
R	v4.5.0	Statistical computing environment for all analyses and figure generation
Reference genome	GRCm39/mm39	Mus musculus reference genome assembly; Genome Reference Consortium Mouse Build 39

Appendix D. Additional QC

Supplementary data files are available in the GitHub repository at:
https://github.com/audreygoddard04/RHAMM-HiFi-lung-metastasis

1. **MultiQC Reports:** Full interactive MultiQC HTML reports aggregating NanoPlot, samtools flagstat, samtools stats, and mosdepth output for all eight samples are available under /supplementary/multiqc_reports/.

2. **Coverage Plots:** Per-chromosome and genome-wide coverage depth plots are available under /figures/coverage_plots/. These plots confirm uniform coverage across autosomes and expected reduced coverage at centromeric and repetitive regions.

3. **Alignment Statistics:** Per-sample alignment statistics from samtools flagstat and samtools stats are summarized in Tables 4 and 5. Full output files are available under /supplementary/alignment_stats/.

4. **NanoPlot HTML Reports:** Interactive NanoPlot HTML summaries per sample are available under /supplementary/nanoplot/. These reports include read length histograms, quality score distributions, and read length vs. quality scatter plots.