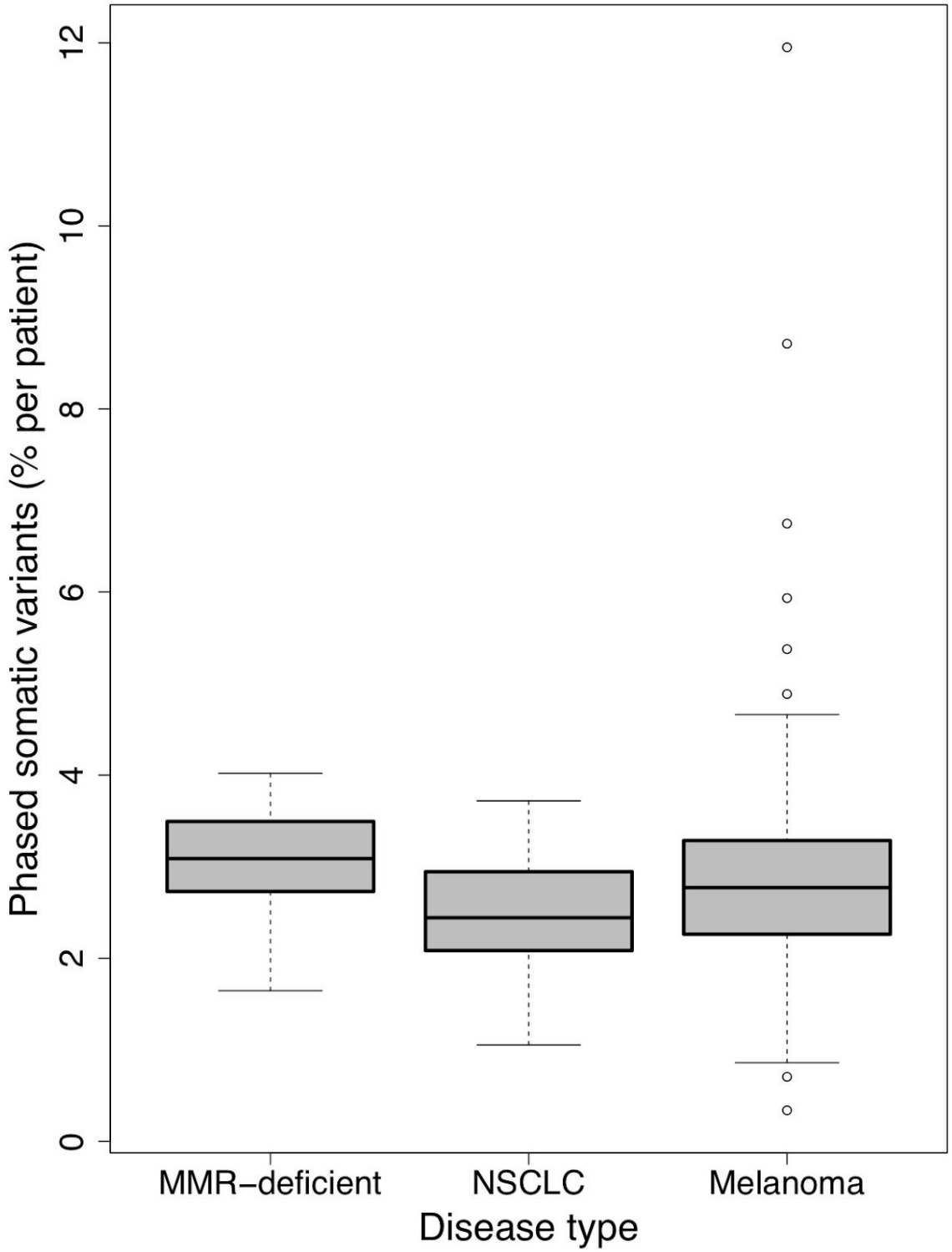
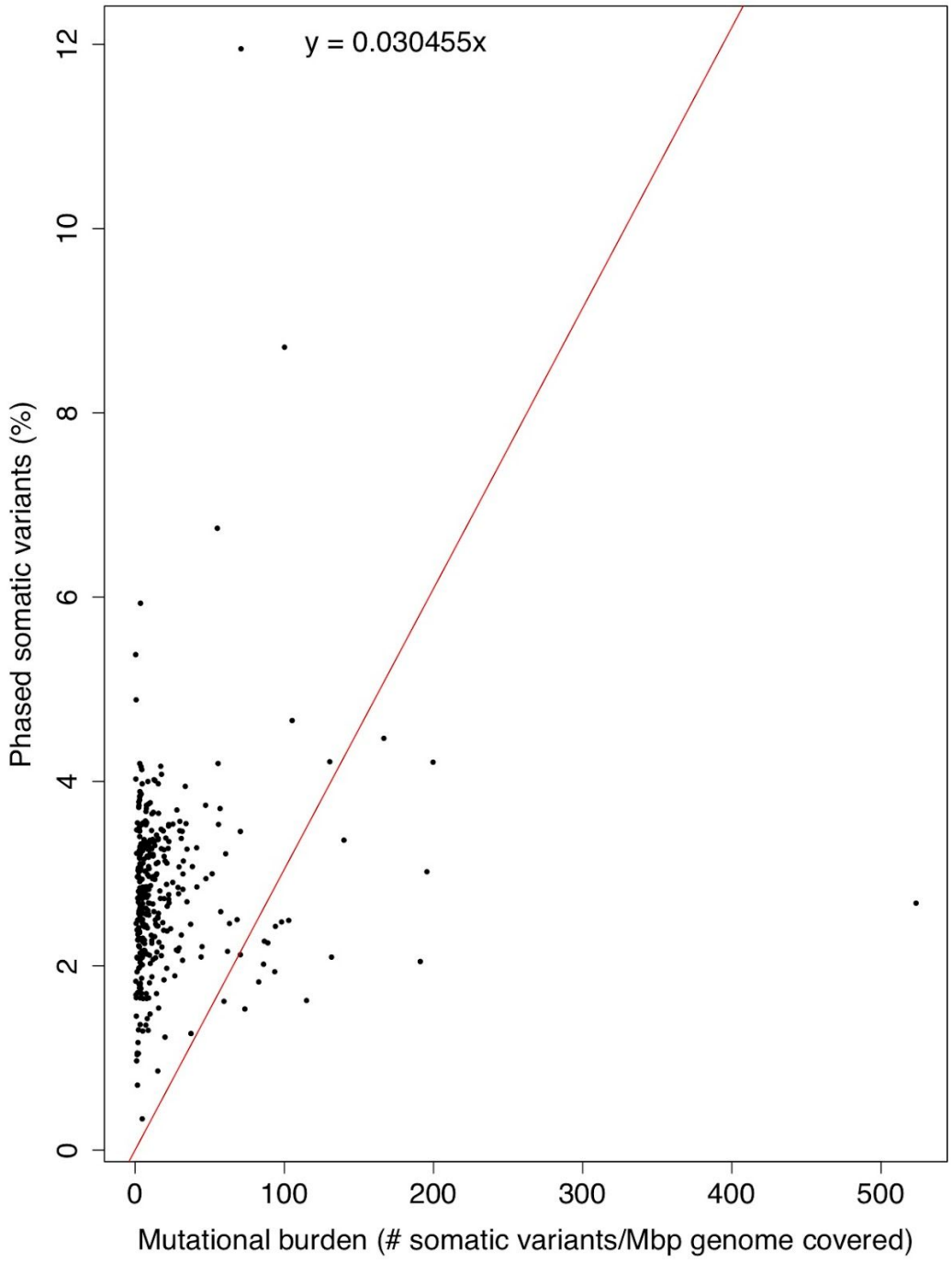
SUPPLEMENTARY FILE 1

Cancer type	Number of patients	Number of tumor samples	Reference
Melanoma	15	15	Amaria et al.
Melanoma	3	7	Carreno et al.*
Melanoma	16	16	Gao et al.
Melanoma	38 (27)	39 (28)	Hugo et al.*
Colon, endometrial, thyroid	28	29	Le et al.
NSCLC	34	34	Rizvi et al.
Melanoma	35	53	Roh et al.
Melanoma	64 (20)	64 (20)	Snyder et al.*
Melanoma	110 (40)	110 (40)	Van Allen et al.*
Melanoma	4	9	Zaretsky et al.

Supplementary Table 1: Summary of patients samples used to identify trends in variant phasing. Publicly available WES data from 10 studies was used to determine the frequency with which somatic variants are phased with germline or other somatic variants (see Materials and Methods). We summarize the study which produced each data set, the cancer types represented, and the number of patients/tumor samples sequenced. Studies that had complementary RNA sequencing reads available for at least a subset of patients are indicated by an asterisk in the “Reference” column, and the number of samples with complementary RNA sequencing data are indicated in parentheses in the “Number of patients” and “Number of tumor samples” columns if different than the number of samples with WES.



Supplementary Figure 1. Co-occurrence of somatic variants by disease type. Box plots demonstrate the per-patient percentage of somatic variants (y-axis) across all tumors that co-occur with either germline variants or other somatic variants across 285 melanoma patients, 34 NSCLC patients, and 28 colon, endometrial, and thyroid cancer patients.



Supplementary Figure 2. Relationship between mutational burden and phasing of somatic variants. The X-axis shows the coverage-adjusted mutation burden per patient (see Methods), while the Y-axis shows the percentage of somatic variants per patient that co-occur with a germline or another somatic variant within 72 bp. The best fit line through the origin is shown in red, with the equation describing it in the top right corner.

Descriptions of Supplementary Tables 2-8 (see Supplementary Files 2 and 3 for data):

Supplementary Table 2. Summary of variant phasing within 33bp. This table summarizes the number of somatic/germline variants per patient and the number of instances of variant phasing as predicted by HapCUT2 analysis of whole exome sequencing data. The “Patient” column lists the patient identifier and the “Tumor_ID” column lists the tumor sample identifier(s); for cases where there is more than one tumor identifier listed, values in the subsequent columns represent a median value across all tumor samples for the patient. The “Disease” column gives the cancer type of the patient. “Total_somatic_mutations” and “Total_germline_mutations” list the number of total somatic and germline mutations for the patient, respectively. “Phased_germline” gives the number of somatic variants phased with at least one germline mutation within 33bp (genomic coordinates), “Phased_somatic” gives the number of somatic variants phased with at least one other somatic mutation within 33bp (genomic coordinates), and “Combined_phasing” gives the number of somatic variants phased with at least one other variant (either somatic or germline) within 33bp (genomic coordinates). “Phased_germline_transcriptomic_coordinates” gives the number of somatic variants phased with at least one germline mutation within 33bp (transcriptomic coordinates), “Phased_somatic_transcriptomic_coordinates” gives the number of somatic variants phased with at least one other somatic mutation within 33bp (transcriptomic coordinates), and “Combined_phasing_transcriptomic_coordinates” gives the number of somatic variants phased with at least one other variant (either somatic or germline) within 33bp (transcriptomic coordinates).

Supplementary Table 3. Summary of variant phasing within 72bp. This table summarizes the number of somatic/germline variants per patient and the number of instances of variant phasing as predicted by HapCUT2 analysis of whole exome sequencing data. The “Patient” column lists the patient identifier and the “Tumor_ID” column lists the tumor sample identifier(s); for cases where there is more than one tumor identifier listed, values in the subsequent columns represent a median value across all tumor samples for the patient. The “Disease” column gives the cancer type of the patient. “Total_somatic_mutations” and “Total_germline_mutations” list the number of total somatic and germline mutations for the patient, respectively. “Phased_germline” gives the number of somatic variants phased with at least one germline mutation within 72bp (genomic coordinates), “Phased_somatic” gives the number of somatic variants phased with at least one other somatic mutation within 72bp (genomic coordinates), and “Combined_phasing” gives the number of somatic variants phased with at least one other variant (either somatic or germline) within 72bp (genomic coordinates). “Phased_germline_transcriptomic_coordinates” gives the number of somatic variants phased with at least one germline mutation within 72bp (transcriptomic coordinates), “Phased_somatic_transcriptomic_coordinates” gives the number of somatic variants phased with at least one other somatic mutation within 72bp (transcriptomic coordinates), and “Combined_phasing_transcriptomic_coordinates” gives the number of somatic variants phased with at least one other variant (either somatic or germline) within 72bp (transcriptomic coordinates).

Supplementary Table 4. Summary of variant phasing within 94bp. This table summarizes the number of somatic/germline variants per patient and the number of instances of variant phasing as predicted by HapCUT2 analysis of whole exome sequencing data. The “Patient” column lists the patient identifier and the “Tumor_ID” column lists the tumor sample identifier(s); for cases where there is more than one tumor identifier listed, values in the subsequent columns represent a median value across all tumor samples for the patient. The “Disease” column gives the cancer type of the patient. “Total_somatic_mutations” and “Total_germline_mutations” list the number of total somatic and germline mutations for the patient, respectively. “Frameshift_phasing_transcriptomic_coordinates” gives the number of cases of either a somatic frameshift variant being phased with a downstream variant (somatic or germline) within 94bp (transcriptomic coordinates), or a frameshift germline variant being phased with a downstream somatic variant within 94bp (transcriptomic coordinates). “Nonstop_phasing_transcriptomic_coordinates” gives the number of cases of either a somatic nonstop variant being phased with a downstream variant (somatic or germline) within 94bp (transcriptomic coordinates), or a nonstop germline variant being phased with a downstream somatic variant within 94bp (transcriptomic coordinates).

Supplementary Table 5. Summary of variant phasing support within RNA. This table summarizes the level of support for variant phasing predicted by HapCUT2 analysis of whole exome sequencing (WES) data within matched RNA sequencing (RNA-seq) data. The “Patient” column lists the patient identifier and the “Tumor_ID” column lists the tumor sample identifier(s); for cases where there is more than one tumor identifier listed, values in the subsequent columns represent a median value across all tumor samples for the patient. “Germline_phased_pairs” gives the number of somatic-germline phased variant pairs predicted by WES, and “Somatic_phased_pairs” gives the number of somatic-somatic phased variant paired predicted by WES (both of these values may reflect redundancy of variants, i.e. one somatic variant may be phased with multiple other variants, each counting as a pair). “Covered_germline_pairs” and “Covered_somatic_pairs” indicate the number of WES-predicted somatic-germline and somatic-somatic phased pairs, respectively, where the positions of both variants are covered by at least one RNA-seq read. “Supported_germline_phasing” and “Supported_somatic_phasing” indicate the number of WES-predicted somatic-germline and somatic-somatic phased pairs, respectively, where the phasing of the two variants is supported by at least one RNA-seq read. “Unsupported_germline_phasing” and “Unsupported_somatic_phasing” indicate the number of WES-predicted somatic-germline and somatic-somatic phased pairs, respectively, where the positions of both variants are covered by at least one RNA-seq read, but phasing is not supported by the read(s). “Novel_germline_phasing” and “Novel_somatic_phasing” indicate the number of instances where a somatic variant is phased with a germline or somatic variant, respectively, in at least one RNA-seq read, and the phasing of those variants was not predicted by WES. “Covered_across_exon_pairs” gives the number of instances of the positions of a somatic variant and another variant (somatic or germline) that are in separate exons being covered by the same RNA-seq read(s), whether or not these variants were predicted to be phased by WES. “Supported_across_exon_pairs” gives the number of instances of a somatic variant and another variant (somatic or germline) that are in separate exons being phased in at

least one RNA-seq read, whether or not these variants were predicted to be phased by WES. "Novel_across_exon_pairs" gives the number of instances of a somatic variant and another variant (somatic or germline) that are in separate exons being phased in at least one RNA-seq read when these variants were not predicted to be phased by WES. "Not_supported_across_exons" gives the number of instances of the positions of a somatic variant and another variant (somatic or germline) that are in separate exons and predicted to be phased by WES being covered by the same RNA-seq read(s), but not having their phasing supported by this RNA-seq data.

Supplementary Table 6. Wallclock times from benchmarking neoepitope calling pipelines. This table summarizes the time in seconds for each process to run in the benchmarked neoepitope calling pipelines for each patient, as well as the average times across patients (in both seconds and minutes). The "Alignment" section covers the times to align tumor and normal samples to the genome using BWA, and the "BAM processing" section covers the times to process the resulting BAM files to prepare for somatic variant calling (marking duplicate reads, sorting BAMs, and performing base quality score recalibration with GATK). The "Somatic variant calling" section covers times to call somatic variants with MuTect and filter those variants, and the "Germline variant calling" section covers times to call germline variants using GATK's HaplotypeCaller and filter those variants. The "HLA typing" section covers the times to perform *in silico* prediction of patient HLA types using Optitype. The "VEP (w/ phasing for pVACseq)" section covers the times to perform phasing with GATK's ReadBackedPhasing for the pVACseq pipeline and annotate the resulting phased VCFs with VEP, while the "VEP (for TSNAD)" section covers the times to annotated unphased VCFs with VEP for the TSNAD pipeline. The "Phasing (HapCUT2)" section covers the times to phase germline and somatic variants with HapCUT2 for the `neoepiscope` pipeline. The "MuPeXI", "pVACseq", "neoepiscope", "NeoPredPipe", and "TSNAD" sections cover the times to perform neoepitope prediction with the specified tools. For pVACseq and TSNAD, times are given both including and excluding the time to run VEP (as a method of comparison with MuPeXI, which runs VEP from within the software; NeoPredPipe, which runs ANNOVAR from within the software; and `neoepiscope`, which performs custom variant effect prediction within the software). The "Full pipelines" section covers the total times to get from raw fastq files to predicted neoepitopes using the required pipelines of the specified tools. See the Methods section in the main text for additional details on the benchmarking steps.

Supplementary Table 7. Summary of neoepitopes predicted during benchmarking of the neoepitope calling pipelines. This table summarizes the neoepitope sequences predicted by one or more neoepitope calling tools during the benchmarking process. Each benchmarked tool (pVACseq, MuPeXI, `neoepiscope`, NeoPredPipe, and TSNAD) is represented by a column. In each column, a 0 indicates that this tool did not predict the given neoepitope sequence, while a 1 indicates that this tool did predict the sequence. For example, a row with a 1 in each column indicates a neoepitope that was unanimously predicted across all tools, while a row with a 1 in only the neoepiscope column indicates that this neoepitope was only predicted by `neoepiscope`.

Supplementary Table 8. Summary of VAF and TCGA transcript expression for neoepitopes predicted only by `neoepiscope` during the benchmarking of neoepitope calling pipelines. For neoepitopes predicted uniquely by neoepiscope during benchmarking, the variant allele frequency (VAF) of the variant of origin(s) is given; if a neoepitope was the result of phased variants with different VAF values, both VAF values are listed. The transcripts of origin for each neoepitope is given. As a surrogate for patient-specific expression data, expression data from TCGA melanoma (SKCM) patients (available at https://osf.io/gqrz9/) was used to determine whether the neoepitopes uniquely predicted by `neoepiscope` are from transcripts commonly expressed in melanoma. A transcript was considered "expressed" in melanoma if the 75th quantile TPM value for that transcript across all patients was greater than 1 TPM. A binary (0/1 for no/yes) expression score is given to indicate whether each neoepitope had at least one possible transcript of origin "expressed" in SKCM patients.