nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-019-0175-1

Use of synthetic DNA spike-in controls (sequins) for human genome sequencing

James Blackburn[1,2,4], Ted Wong[1,4], Bindu Swapna Madala[1], Chris Barker[1], Simon A. Hardwick[1,2], Andre L. M. Reis[1,2], Ira W. Deveson[1,2*] and Tim R. Mercer[1,2,3*]

Next-generation sequencing (NGS) has been widely adopted to identify genetic variants and investigate their association with disease. However, the analysis of sequencing data remains challenging because of the complexity of human genetic variation and confounding errors introduced during library preparation, sequencing and analysis. We have developed a set of synthetic DNA spike-ins—termed 'sequins' (sequencing spike-ins)—that are directly added to DNA samples before library preparation. Sequins can be used to measure technical biases and to act as internal quantitative and qualitative controls throughout the sequencing workflow. This step-by-step protocol explains the use of sequins for both whole-genome and targeted sequencing of the human genome. This includes instructions regarding the dilution and addition of sequins to human DNA samples, followed by the bioinformatic steps required to separate sequin- and sample-derived sequencing reads and to evaluate the diagnostic performance of the assay. These practical guidelines are accompanied by a broader discussion of the conceptual and statistical principles that underpin the design of sequin standards. This protocol is suitable for users with standard laboratory and bioinformatic experience. The laboratory steps require ~1–4 d and the bioinformatic steps (which can be performed with the provided example data files) take an additional day.

Introduction

NGS can be used to identify genetic variation and diagnose disease-associated mutations, and has become a principal tool in biomedical research and clinical diagnostics[1]. However, numerous variables influence the sensitivity and precision of variant detection using NGS. Sequencing depth and read length are key variables for any NGS assay[1,2], and sequencing errors and PCR amplification biases introduced during library preparation further impact performance[3–8]. During downstream bioinformatic analysis, outcomes may vary between alternative software tools and parameter settings, and incorrect analytical assumptions or user errors have the potential to cause misdiagnosis[9]. These and other variables accumulate across an NGS workflow, confounding the accurate detection of genetic variants.

To address this issue, we have developed spike-in DNA standards, termed 'sequins'. Sequins are synthetic mirror-image representations of human DNA sequences, including instances of genetic variation, that act as internal reference standards during NGS experiments[10–12] (Fig. 1). Sequins can be added to a user's DNA sample before library preparation and sequencing. Owing to the artificial nature of sequin sequences, reads derived from sequin standards can be distinguished from human DNA sequences (or other natural sequences) in the resulting library. This allows sequins to be analyzed as internal controls that do not interfere with the analysis of the accompanying sample[10–12] (Box 1; Fig. 2).

Sequins are subject to the same technical variables as the accompanying DNA sample and, accordingly, can be used to assess the impact of laboratory and bioinformatic variables at any stage of the NGS workflow. Sequins can measure the diagnostic performance (e.g., sensitivity and precision) of a given NGS assay, enable rapid troubleshooting and operational quality control, and act as scaling factors by which to normalize between multiple samples[10–12].

In this protocol, we describe the use of sequins as internal reference standards during human whole-genome sequencing (WGS) and targeted sequencing of cancer genes in tumor

[1] Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, Australia. [2] St Vincent's Clinical School, Faculty of Medicine, UNSW Australia, Sydney, Australia. [3] Altius Institute for Biomedical Sciences, Seattle, WA, USA. [4] These authors contributed equally: J. Blackburn, T. Wong. *e-mail: i.deveson@garvan.org.au; t.mercer@garvan.org.au

Box 1 | What are sequins?

DNA sequences have an inherent 5′–3′ directionality that is observed in all cellular processes, including DNA replication, transcription and translation. By arranging any human DNA sequence in reverse nucleotide order, we can generate a synthetic DNA sequence that directly mirrors the original human sequence (Fig. 2a). For example, we can reverse the sequence 5′-GACTGA-3′ to form the mirrored sequence 5′-AGTCAG-3′, which represents a distinct DNA molecule (note that we have reversed the original sequence directly, not taken its reverse-complement)[10].

Although the mirrored DNA sequence is distinct, it retains the same nucleotide composition and repetitiveness of the original human sequence. Owing to their shared properties, the mirrored sequence performs equivalently to the human sequence during laboratory (e.g., PCR amplification, hybridization and sequencing) and bioinformatic (e.g., alignment and variant detection, with respect to a mirrored reference sequence) processes. Therefore, the mirrored sequence constitutes a good proxy for its corresponding human DNA sequence[10].

This design principle can be applied to generate synthetic mirrored sequences that directly represent almost any feature of the human genome, including reference gene sequences or instances of genetic variation. In principle, we can represent genetic variants of any type, size and context, including single nucleotide variants (SNVs), small insertions/deletions (indels) and large structural or copy-number variants. Common and clinically relevant variants such as cancer driver mutations (Fig. 2a) can be represented, as well as variants occupying analytically challenging regions of the genome, such as microsatellite repeats (Fig. 2b).

Synthetic mirrored sequences are manufactured into small DNA molecules (typically, 2–10 kb in length) that we call ‘sequins’ (sequencing spike-ins). Hundreds of individual sequins are typically combined to formulate a synthetic mixture, together representing a diverse array of genetic features. By combining sequins in precise stoichiometric ratios, we can emulate quantitative features of genome biology such as variant allele frequencies or copy-number variation (Fig. 2c,d). By staggering the concentrations of multiple sequins, we create internal reference ladders that span the relevant dynamic range for a given genetic feature. For example, adding a variant sequin by itself establishes a homozygous genotype, whereas combining variant and wild-type sequins at equal concentrations establishes a heterozygous genotype. Further dilution of variant sequins relative to their corresponding wild-type sequins establishes an allele frequency reference ladder that measures the sensitivity and quantitative accuracy with which somatic mutations are detected in a given sample (Fig. 2c)[10].

(and matched normal) samples, with sequins being used to evaluate the detection of germline and somatic variants, respectively.

Overview of the procedure

In this protocol, we provide step-by-step instructions for the use and analysis of sequins in conjunction with (i) WGS of a reference human genome sample (genomic DNA (gDNA) from individual NA12878 (ref. [13])) and (ii) targeted sequencing of a mock ‘tumor’ and matched ‘normal’ sample pair comprising a mixture of common cancer cell lines[14]. There is some overlap between these procedures and, for brevity, we have not repeated steps that are shared between them. Nevertheless, there are several important distinctions between these two approaches, with additional laboratory steps and equipment setup required for targeted sequencing and different analytical approaches used to identify germline and somatic mutations.

Detailed instructions are provided for the use of sequins in the laboratory, including how sequins should be diluted, added to DNA samples and prepared for sequencing (Steps 1–11). We also provide instructions for the bioinformatic analysis of the resulting libraries (Steps 12–25), including the use of popular tools for NGS analysis (such as BWA[15], SAMtools[16] and Strelka2[17]), as well as the anaquin toolkit[18], which we have purpose-built for sequin analysis. We first focus on the identification of germline variants (single-nucleotide variants (SNVs) and indels) in the NA12878 sample analyzed by WGS (Steps 12–21). We then evaluate the detection of somatic mutations in the tumor/normal sample pair analyzed by targeted sequencing (Steps 22–25).

Raw NGS libraries that we have generated via the laboratory steps described here are available for download (Equipment), allowing users who have not prepared their own libraries to complete the bioinformatic protocol (Steps 12–25).

Use of sequins with WGS applications

Using sequins with WGS is relatively straightforward; the fraction of reads in the output library derived from sequins should reflect the fractional abundance at which sequins were originally added to a user’s DNA sample. Accordingly, sequins should be added at sufficient abundance (~1–2%) to achieve matched sequencing coverage with the accompanying diploid human genome while minimizing the fraction of reads sacrificed to spike-in standards. A guide to incorporating the correct amount of sequins for WGS analysis can be found in Table 1. The combined sample (user DNA sample and sequins) is then processed and sequenced via conventional methods.

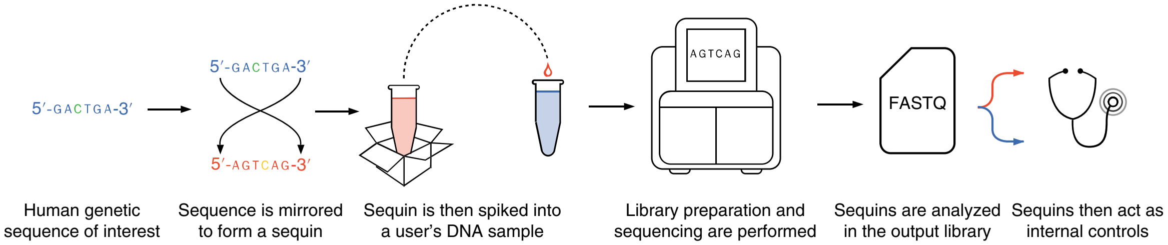

5′-GACTGA-3′
5′-GACTGA-3′
5′-AGTCAG-3′
AGTCAG
FASTQ
Human genetic sequence of interest
Sequence is mirrored to form a sequin
Sequin is then spiked into a user's DNA sample
Library preparation and sequencing are performed
Sequins are analyzed in the output library
Sequins then act as internal controls

Fig. 1 | Schematic showing the design and use of sequins in NGS experiments. Any human genetic sequence can be mirrored to create a synthetic DNA standard (a 'sequin'). Sequins can be added to a human DNA sample before performing library preparation and sequencing. Owing to their synthetic sequence, sequin-derived reads can be distinguished from sample-derived reads in the resulting library, allowing sequins to be analyzed as internal controls.

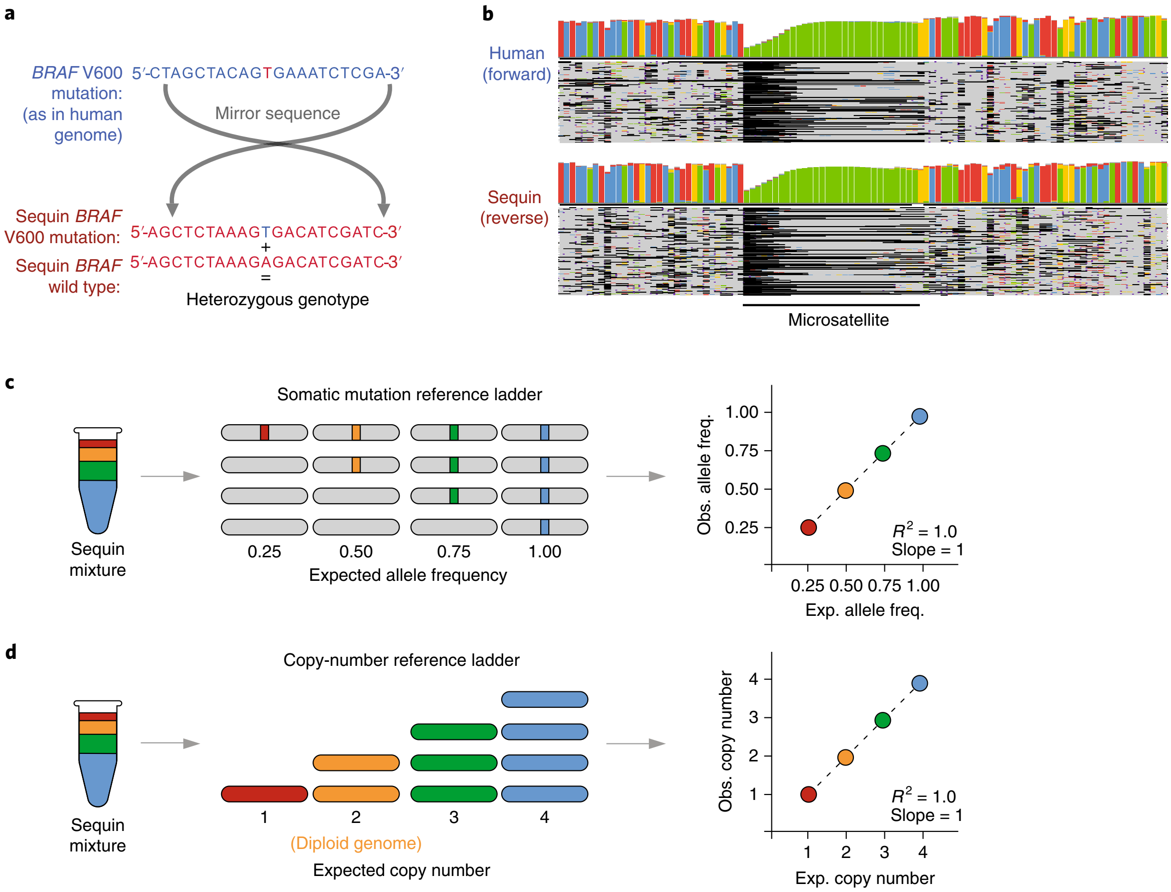


Fig. 2 | Sequin design principles. a, The *BRAF* gene sequence, containing the cancer driver mutation V600E, from the human genome is reversed to form a mirrored sequin sequence that acts as a matched control. **b**, Genome browser view showing alignments (Oxford Nanopore reads) at a human microsatellite sequence (upper) and corresponding sequin (lower). Insertion/deletion and mismatch sequencing errors present in human alignments are recapitulated in the sequin sequence. **c**, Quantitative allele frequency ladders can be established by sequentially titrating sequins representing variant (colored) and reference (gray) allele sequences. **d**, Reference ladders for measuring copy-number variation can also be formed by combining sequins at staggered concentrations. See Box 1 for a discussion of sequin design principles.

Use of sequins with targeted sequencing applications

During targeted NGS approaches, gene sequences of interest are typically enriched by hybridization to complementary biotinylated oligonucleotide probes (capture) or amplified by multiplexed PCR (amplicon). This confers greater sequencing coverage, thereby improving sensitivity for variant detection and allowing numerous samples to be multiplexed at reduced cost[19,20]. However, the target-enrichment process is a major additional source of technical variation affecting diagnostic performance[7]. Sequins are compatible with both targeted sequencing strategies, in which they provide the additional benefit of evaluating biases introduced during target enrichment.

Table 1 | Guidelines for diluting genome sequins according to sample DNA amounts/library preparation method

Application	Sample DNA amount	Input genome sequin amount	Genome sequin volume to add (dilution)
WGS	1 ng	0.02 ng	1 µl (1:500)
WGS	10 ng	0.2 ng	1 µl (1:50)
WGS	100 ng	2.0 ng	1 µl (1:5)
WGS	200 ng	4.0 ng	1 µl (2:5)
WGS	1,000 ng	20.0 ng	2 µl (neat)
Targeted	100 ng	0.5 ng	1 µl (1:20)
Targeted	200 ng	1.0 ng	1 µl (1:10)
Targeted	1,000 ng	5.0 ng	1 µl (1:2)

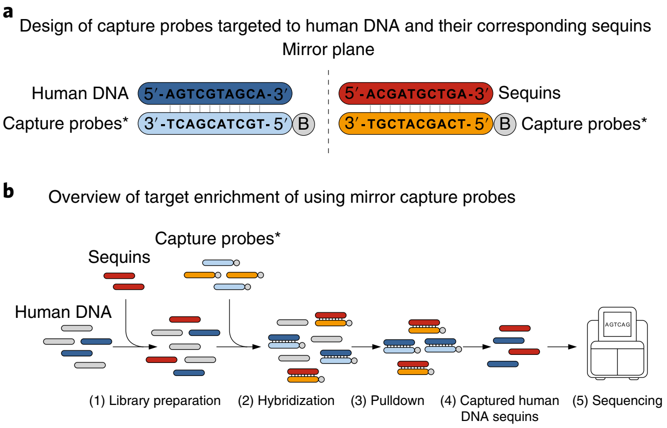


Fig. 3 | Compatibility of sequins with targeted sequencing. a, Diagram showing the design of sequins (red) and their corresponding capture probes (orange) to mirror genome regions of interest (dark blue) and their corresponding capture probes (light blue). **b**, Schematic showing the addition of sequins to a human DNA sample before library preparation, target enrichment and sequencing. *Complementary biotinylated oligonucleotides.

To use sequins in a targeted sequencing experiment based on complementary probe hybridization (e.g., exome sequencing), additional probes must be designed to allow for the concurrent enrichment of sequins alongside target genes in the accompanying sample (Fig. 3). Given that sequin and sample DNA sequences do not hybridize to, and therefore compete for, the same capture probes, the capture of sequins does not impact the capture of endogenous sequences[10].

Ideally, the additional capture probes that target sequins should perfectly mirror the sequences of the probes that target the corresponding human genes (Fig. 3). The enrichment reaction, by which probes hybridize and pull down human genome sequences, is thus mirrored by the enrichment reaction for the corresponding sequin controls. By mirroring this reaction, sequins can evaluate the hybridization and capture of matched gene sequences within the same reaction conditions. Therefore, for custom gene panels, we recommend that users design probes to target sequins that directly mirror their own custom gene probes.

Sequins are also compatible with multiplexed PCR strategies that amplify genome sequences of interest before sequencing. Integrating sequins with these approaches requires additional primers to amplify sequins along with endogenous human genome sequences. As with capture probes, the primers used to amplify sequins should ideally mirror the primers used to amplify corresponding regions of the human genome. This allows sequins to mirror and evaluate the amplification of endogenous human sequences while avoiding primer sequence competition and dimerization.

When used with targeted sequencing approaches, care must be taken not to add sequins to the sample in excessive stoichiometry, because exponential amplification can then lead sequins to overpopulate the sequenced library. Within this protocol, we provide guidelines for the amount of sequins to add to a human DNA sample (Table 1). However, we recommend that users working with custom capture/amplicon panels independently validate the suggested spike-in amounts, as these values may need to be adjusted depending on their panel design and library preparation protocol.

The prevalence of sequins in a target-enriched library preparation can also be affected by additional variables such as DNA sample quality, library complexity and PCR amplification. To mitigate risk, we recommend that users validate library composition using qPCR before sequencing. Further information on estimation of the dilution factor and validation of libraries with qPCR is provided below.

Other applications

Sequins are simply DNA molecules that are added to a DNA sample of interest and are therefore compatible with most library preparation methods and sequencing technologies. We have validated the use of sequins with a wide range of popular library preparation kits, short- and long-read sequencing technologies and a variety of bioinformatic software, as mentioned in Box 2 (indeed, sequins constitute an ideal method for benchmarking different reagents or methods). We have also tested sequins in comparison with a range different sample types, including fresh, frozen and formalin-fixed, paraffin-embedded (FFPE) samples.

During the bioinformatic protocol, we show how sequins can be used to evaluate the detection of germline and somatic variants. However, there is almost no limit to the analyses that can be performed with sequins, and we encourage users to perform tailored analyses according to their own interests or requirements (Box 2). Similarly, although we have designed sequins for use with human genome sequencing, sequins are also compatible (and similarly useful) with samples from other organisms.

We have also designed a range of different sequin mixtures for alternative NGS applications. For example, we have developed RNA sequins for transcriptomic studies using RNA sequencing (RNAseq)[21], as well as sequins for use with metagenome sequencing experiments[22]. Additional details on the design and use of sequins (as well as further supporting information, including software, protocols and tutorials) are available at http://www.sequinstandards.com.

Advantages

As internal reference standards, sequins provide multiple advantages that can be broadly summarized as shown in the following sections.

Diagnostic performance for each library

The sensitivity of an NGS assay can be assessed by measuring the true-positive and false-negative rate of detection of synthetic sequin variants. We can also measure the rate at which false-positive variants are erroneously detected within sequin sequences and thereby determine the precision (or specificity) of the assay. These metrics together enable the diagnostic power of each individual NGS library to be calculated.

Sequins can also be used to infer the uncertainty associated with measurements of quantitative features, such as allele frequency, enabling the quantitative accuracy of a given library to be evaluated. In addition, sequins can be used to assess the performance of sequencing and read alignment within corresponding regions of the human genome in order to identify analytical blind spots.

Operational performance, quality control and troubleshooting

Sequins can be used to evaluate the performance of each step within the NGS workflow, ensure that minimum performance thresholds are met, and provide oversight of laboratory and bioinformatic operations. For example, the routine measurement of sequins can quickly identify decreases in operational performance and enable rapid troubleshooting to identify the causative problem. Sequins can also assess the impact when new equipment, reagents or personnel are introduced into the NGS workflow. These advantages are particularly pronounced when surveillance extends over large sample cohorts or across long time periods.

Normalization

Sequins can measure and mitigate the technical variation that a sample accrues as it proceeds through the NGS workflow and can thereby enable normalization between multiple samples. As a constant set of DNA molecules that are spiked into multiple samples, sequins can be used to calculate normalization scaling factors. Notably, the wide quantitative range encompassed by sequins enables

Box 2 | What is in the human genome sequin mixture?

Within this protocol, we largely focus on common variants and somatic mutations. However, there is a diversity of additional sequins within the mixture that we encourage users to explore. Here, we provide a brief description of the different features represented in our human genome–sequin mixture.

Common genetic variations
Number of sequins = 101; sequin naming code = GV (germline variant); genotypes represented = homo/heterozygous; applications = WGS
Any individual harbors many common genetic variants that are widely distributed throughout the population. We have developed a set of sequins representing homo/heterozygous SNVs and indels that constitute a cross-section of the common genetic variation found in an individual genome. These provide a reference with which to assess the performance of standard germline variant detection using WGS in conjunction with your choice of downstream bioinformatic tools (e.g., GATK[26]).

Difficult genetic variations
Number of sequins = 67; sequin naming code = DV (difficult variant); genotypes represented = heterozygous; applications = WGS
Low-complexity or repetitive sequences are among the most polymorphic sites in the human genome, and variants at these sites have established roles in a range of diseases. However, it can be difficult to distinguish small variants from sequencing or alignment errors at repetitive sites. Sequences that are especially GC- or AT-rich also hinder NGS analysis. Therefore, we have developed sequins that represent germline variants occurring at simple repeats (mono-, di-, tri- and quad-nucleotide) or within GC/AT-rich regions. These can be used to evaluate the ability of a given NGS assay to identify variation in challenging regions of the human genome and to gauge the strengths and weaknesses of alternative protocols (e.g., standard versus PCR-free library preparations).

Haplotypes
Number of sequins = 47; sequin naming code = HP (haplotype); genotypes represented = homo/heterozygous; applications = WGS, long-read sequencing
We have developed a set of large sequins (~6 kb) that represent a pair of paternal/maternal haplotype sequences located on each human chromosome (paternal-only on the Y chromosome). Each standard encodes multiple SNVs and indels that are either shared (homozygous) or unique to the maternal/paternal haplotype (heterozygous). Because the physical relationships of encoded variants are known, these sequins can be used to evaluate the performance of bioinformatic tools for variant discovery and phasing, or the accuracy of de novo sequence assembly. Moreover, their size makes these sequins amendable to analysis by single-molecule (e.g., PacBio[27] or Oxford Nanopore[28]) or synthetic long-read sequencing approaches (e.g., 10× Genomics[29]), enabling benchmarking of these techniques.

Structural variants
Number of sequins = 58; sequin naming code = SV (structural variant), TL (translocation), PV (HPV insertion); genotypes represented = heterozygous; applications = WGS, long-read sequencing
Structural variants (SVs) are a major form of human genomic variation and have recurrent roles in inherited diseases and cancer. However, the size and diversity of SVs, and their common inclusion of repetitive sequences pose challenges for SV detection using NGS. We have developed a set of sequins that represent a broad selection of SV types, including large deletions, inversions and tandem duplications, chromosomal translocations and insertions of both exogenous viral sequences and mobile elements. This set includes both common variants, which can often be compared to matched examples within an accompanying human DNA sample, and clinically relevant events, such as known oncogenic translocations and viral insertions. For each synthetic SV, the non-affected allele is also represented, thereby emulating a heterozygous genotype.
SV sequins provide an internal reference that can be used to assess the sensitivity for detecting different SV types and sizes, as well as the accuracy with which SV breakpoints are resolved. Sequins can be used to easily evaluate the performance of software tools that identify different types of SVs from different sources of evidence, including read depth (e.g., CNVnator[30]), chimeric reads (e.g., LUMPY[31]) and de novo sequence assembly (e.g., Pamir[32]). This makes SV sequins an ideal resource for benchmarking new approaches, especially those that use emerging technologies such as Oxford Nanopore sequencing (e.g., Sniffles[33]).

Inherited disease genes
Number of sequins = 105; sequin naming code = CL (clinical gene); genotypes represented = reference; applications = WGS, Exome
Heritable mutations in many human genes cause disease. We have developed a set of sequins that represent the clinically informative domains/exons from >90 genes that are associated with heritable human diseases, including cystic fibrosis, hemophilia, cardiac myopathies, hereditary cancer and triplet-expansion disorders, as well as pharmacogenes associated with adverse drug reactions. In each case, we have represented only the human reference sequence, thereby providing an internal standard with which to interpret candidate variants detected in the accompanying human DNA sample. For example, by representing the *HTT* gene, we provide a standard that can be used to assess the reliability of a possible triplet repeat expansion in this gene that was detected with NGS. Given that the majority of relevant regions are exonic, this provides a useful resource, in addition to WGS, for assessing performance during exome sequencing.

Cancer genes
Number of sequins = 141; sequin naming code = CL (clinical gene); genotypes represented = reference; applications = WGS, Exome, Oncology panel
Many human genes have been causatively associated with cancer, and the detection of mutations in these genes can inform patient prognosis and treatment. We have developed a set of sequins that represent the clinically informative domains/exons from >100 genes causally associated with human cancers, such as *BRCA1, TP53, ERBB2* and *ALK*. For each gene, we have represented the wild-type sequence, providing a reference with which to interpret candidate germline and somatic mutations detected in the accompanying human DNA sample. Cancer gene exons can provide internal controls for WGS or exome sequencing, and can be used during the design, validation and ongoing quality control of targeted oncology gene panels.

Somatic mutations
Number of sequins = 144; sequin naming code = CV (cancer variant), CM (multiple linked cancer variants); genotypes represented = somatic; applications = WGS, Exome, Oncology panel
The accurate detection of somatic mutations in tumor DNA samples can inform prognosis and treatment for cancer patients. However, due to tumor evolution and the presence of non-tumor cells within a sample, somatic mutations often occur at heterogeneous (typically low) variant allele frequencies (VAFs). High sequencing coverage is required to detect these low-VAF mutations, which can be difficult to distinguish from sequencing errors. We have developed a set of sequins that represent known cancer driver mutations at staggered concentrations to form a quantitative VAF

Box 2 | What is in the human genome sequin mixture? (Continued)

ladder (100–0.1%). This ladder provides an internal scale that can be used to evaluate the sensitivity, precision and quantitative accuracy with which somatic mutations can be detected in a given NGS library. To ensure compatibility with popular bioinformatic tools (e.g., Strelka2[17], Mutect2[34]) that call somatic mutations in tumor DNA by comparison to a matched-normal sample, we provide a separate 'normal' sequin mixture, in which only the wild-type gene sequence at each cancer mutation is represented, thereby providing a background against which somatic mutations can be called.

Clinical microsatellites
Number of sequins = 12; sequin naming code = MS (microsatellite); genotypes represented = stable/unstable; applications = WGS, Exome, Oncology panel
Microsatellite instability (MSI) is indicative of mismatch repair deficiency in cancer, informing patient prognosis and treatment decisions. MSI diagnosis involves the detection of insertions and deletions at microsatellite sequences (short tandem repeats) throughout the genome. However, repeats are refractory to NGS analysis and prone to sequencing and alignment errors that confound MSI diagnosis. To evaluate the diagnosis of MSI with WGS or targeted sequencing approaches, we have designed sequins that represent both stable and unstable instances of microsatellite loci that are commonly used as markers for MSI profiling (Bethesda panel[35]). These allow false-positive and false-negative results to be identified during MSI profiling and can be used to assess the impact of technical variables during library preparation and sequencing (e.g., read-length, number of PCR cycles) on diagnostic performance.

Immune receptors
Number of sequins = 62; sequin naming code = IM (immune receptor); genotypes represented = clonal; applications = WGS, Exome, Immune panel
Sequencing of immunoglobulin and T-cell receptor loci following somatic recombination and hyper-mutation can reveal the immune repertoire within a sample and can indicate the presence of clonal immune-cell populations. However, owing to the number, repetitiveness and complexity of possible clonotypes, immune-repertoire profiling remains challenging. To improve and standardize this technique, we have developed sequins that represent somatically rearranged immunoglobulin (*IgH*, *IgL* and *IgK*) and T-cell receptor genes (*TCRA/D*, *TCRB* and *TRCG*). These can be used to measure the accuracy with which clonotype sequences are determined and can indicate the quantitative accuracy of measurements of clonal cell populations. Immune sequins are compatible with WGS, targeted immune repertoire sequencing, and emerging long-read sequencing techniques that enable single-molecule B- and T-cell receptor characterization.

Human leukocyte antigen alleles
Number of sequins = 5; sequin naming code = HL (HLA gene); genotypes represented = heterozygous; applications = WGS, Exome, HLA panel
The human leukocyte antigen (HLA) genes have established roles in autoimmune disease etiology, adverse drug reactions and cancer development. However, accurate genotyping of HLA genes can be difficult because of high rates of polymorphism and the presence of additional homologous sequences in the genome. Accordingly, we have developed sequins that represent two common human alleles for each of the major HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DR* and *HLA-Q*). These provide internal reference standards for HLA typing by WGS, exome or targeted HLA sequencing approaches and can be used to benchmark HLA-typing software (e.g., *HLAscan*[36]).

Mitochondria
Number of sequins = 3; sequin naming code = MT (mitochondrial genome); genotypes represented = reference; applications = WGS, Mitochondrial sequencing
NGS is emerging as the primary tool for diagnosing mitochondrial mutations that are associated with a range human disease. However, owing to mitochondrial heteroplasmy, mutations appear at heterogeneous (often low) frequencies, making them difficult to detect. We have developed three large sequins to represent the majority of the mitochondrial genome, thereby providing a useful reference with which to interpret candidate mutations detected in the accompanying human mitochondrial DNA. Mitochondrial standards can be analyzed via WGS or targeted mitochondrial sequencing approaches.

local scaling factors to be determined that mitigate the non-linear impact of technical variables such as PCR amplification biases.

Limitations

When using spike-in controls in any NGS experiment, some sequencing reads are necessarily sacrificed, thereby reducing the total pool available for analysis of the sample of interest. This limitation is of minimal importance during WGS analysis, in which no more than ~1–2% of reads must be sacrificed in order to achieve matched stoichiometry between sequins and the diploid human genome. During targeted sequencing approaches, a larger fraction of the sequenced library may be sacrificed, with this depending on the collective size of captured genome regions relative to captured sequin sequences and therefore differing for alternative capture panel designs. We encourage users to incorporate sequins thoughtfully into their designs, targeting only those standards that will provide relevant information for their specific project.

Sequins are compatible with most standard NGS applications and different types of samples (FFPE, fresh, frozen). However, additional caution should be exercised when using sequins with DNA samples of poor or degraded quality, in which case sequins may undergo library construction preferentially at the expense of the accompanying DNA sample. This limitation can be offset by adding sequins at a lower fractional abundance and/or by using qPCR to ensure that sequins do not overpopulate the constructed library (see below).

Alternative methods

Many alternative reference standards for human genome sequencing are available, including natural DNA samples, synthetic sequences and in silico reference datasets, and we recommend their use in establishing a rigorous and accurate NGS workflow[12].

In particular, we advise the use of reference human genome materials in conjunction with sequins to properly assess the veracity of clinical diagnosis[13,23]. gDNA derived from well-characterized cell lines constitutes a cheap and renewable source of reference materials[24]. An advantage of genome reference materials is that they encompass the full size and complexity of the genome and perform similarly to patient DNA (i.e., are commutable)[12]. Many clinical laboratories routinely sequence DNA from the NA12878 cell line (GM12878) as a process control for their NGS workflow. Derived from a healthy female individual, this genome has been extensively characterized using various sequencing technologies, resulting in a high-confidence set of genotypes across the 'high confidence' fraction of the genome[13].

Alternative spike-in controls for human genome sequencing are also available. In the control plasmid spike-in genome (CPSG) system, clinically relevant genetic variants embedded in native human genome sequence are labeled with short molecular barcodes that allow them to be distinguished from accompanying sample DNA[25]. CPSG standards are compatible with most NGS methods, including targeted sequencing, in which they have the advantage of being compatible with existing capture panel designs that target corresponding human genes. However, CPSGs do not encompass the breadth and diversity of genome sequences and genetic variants that are represented by sequins. Because they are composed of natural human DNA sequences, CPSG standards could be used in tandem with sequins, with reads derived from each spike-in set being readily distinguished.

Laboratory procedure

In this protocol, we describe the practical use of sequins within the laboratory, including how to resuspend and dilute sequins before addition to your DNA sample of interest. Please note that in the laboratory procedure, we describe how to use sequins in conjunction with both WGS (Step 2A) and targeted sequencing (Step 2B).

The aim of the WGS protocol is to demonstrate the use of sequins to assess the sensitivity and accuracy with which germline variants are detected. We describe the addition of sequins to NA12878 gDNA, which then undergoes library preparation (using the TruSeq DNA PCR-free Library Preparation Kit) before sequencing (using the HiSeq X Ten).

The aim of the targeted sequencing protocol is to demonstrate the use of sequins to assess the sensitivity and accuracy with which somatic mutations are detected. We use a custom gene panel (SeqCap EZ Developer Enrichment Kit) that captures 408 human cancer genes, as well as corresponding sequin standards. In this capture panel, human and sequin capture probes are designed to be mirrored sequences (Fig. 3). However, users may wish to design their own capture panel and should customize the capture probes that target sequins so that these mirror the corresponding probes targeting their own genes of interest.

Targeted sequencing is performed on a mock tumor sample comprising a mixture of NA12878 (69%), MCF7 (30%) and K562 (1%) cell lines. MCF7 is a well-characterized breast carcinoma cell line, and K562 is an established chronic myelogenous leukemia (CML) cell line[14]. This mock tumor sample is matched with a normal sample that solely constitutes NA12878 gDNA. We have assembled two alternative sequin mixtures; the 'Genome v2 Mix' contains synthetic germline variants as well as somatic mutations, whereas the 'Matched Normal v2 Mix' contains only germline variants. In the example Procedure, the Genome Mix is added to the tumor DNA sample, whereas the Matched Normal Mix is added to the matched normal DNA sample. The Matched Normal sequin mixture provides an unmutated background sequence, against which somatic mutations in the Genome Mix can be identified.

Following this, both samples undergo library preparation (using the KAPA HyperPlus library preparation kit), followed by target enrichment with a custom gene panel array (using the SeqCap EZ Developer Enrichment Kit). We also describe additional validation steps in which qPCR is used to evaluate the scale of target enrichment and the anticipated sequin library fraction.

Bioinformatic procedure

The sequencing of a human DNA sample with added sequins will generate an output FASTQ library file containing reads derived from both the sample DNA and the accompanying sequins. We use the

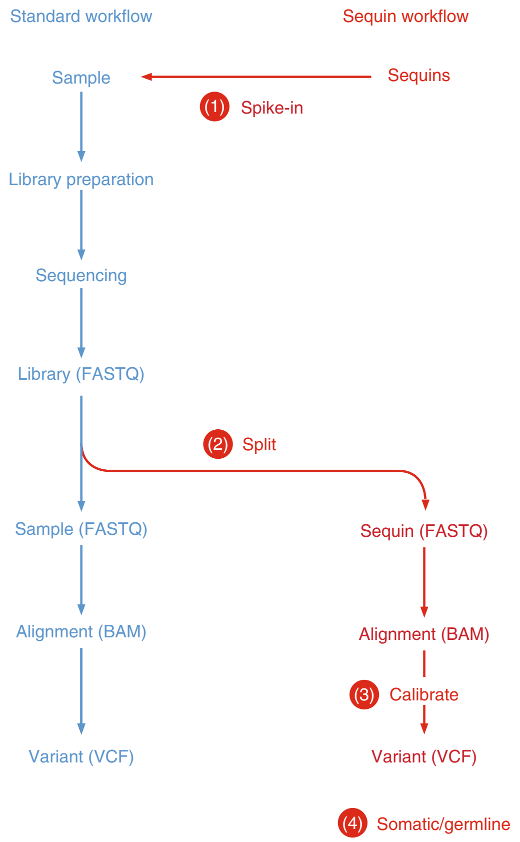


Fig. 4 | Overview of protocol for sequin use in human genome sequencing. (1) Sequins are added to a user's DNA sample before library preparation and sequencing. (2) The resulting FASTQ library is partitioned into sequin-derived and sample-derived reads, using anaquin 'split', which also reverses the orientation of the sequin reads. (3) The partitioned reads are aligned to the human reference genome (using BWA), and sequin reads are down-sampled using anaquin 'calibrate' to achieve coverage matched to the human sample. (4) Sample- and sequin-derived reads are then analyzed in parallel using standard bioinformatic software (Strelka2). Analysis of sequins can evaluate the diagnostic performance of the experiment, assess the accuracy of quantitative measurements and normalize multiple samples.

anaquin 'split' tool to identify and partition sequin-derived reads from sample-derived reads (Fig. 4). During this process, sequin-derived reads are also mirrored so that their sequence orientation matches that of the human reference genome. This enables the analysis of sequin-derived libraries with mainstream bioinformatic tools and ensures compatibility with existing gene, variant and other annotations. Following the 'split' procedure, users will have two library files, one FASTQ file containing sample-derived reads, and one FASTQ file containing sequin-derived reads, both of which can be aligned to the human reference genome (Fig. 4).

Users can process the two FASTQ files in parallel, using their preferred bioinformatic workflow. In this protocol, we first align the two libraries to the human genome (hg38) using BWA[15]. We then use the anaquin 'calibrate' tool to calibrate the depth of sequin alignments to match the sequencing depth of the accompanying sample (Fig. 5). This ensures that sequencing coverage, a key variable in NGS analysis, is matched between sample and sequins before further analysis. We then identify variants from both sequin- and sample-derived alignments using Strelka2[17], with germline variants and somatic mutations detected in the WGS and the targeted sequencing experiments, respectively (Fig. 4).

Sequins can be analyzed to assess performance at each step in the NGS workflow. Within this protocol, we show how the previous variant detection step can be evaluated using the anaquin 'germline' and 'somatic' tools, with both tools producing automated reports that evaluate variant detection in sequins and the accompanying human sample. Similar automated reports can also be generated directly from the raw FASTQ library files, using the anaquin 'split' tool, which performs an alignment-free analysis of library composition. At completion, users will have identified variants in sequins and the accompanying human samples and will have evaluated diagnostic performance of the example assays.

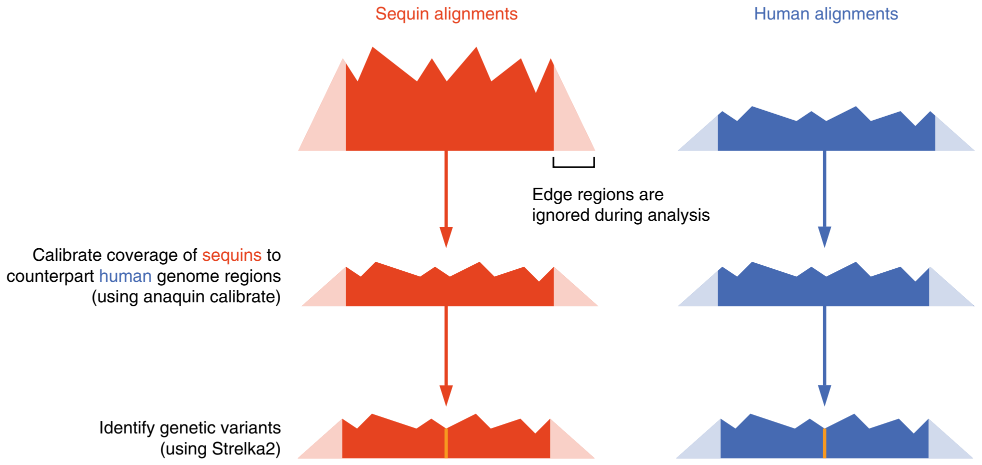

Sequin alignments
Human alignments
Edge regions are ignored during analysis
Calibrate coverage of sequins to counterpart human genome regions (using anaquin calibrate)
Identify genetic variants (using Strelka2)

Fig. 5 | Calibration of sequin coverage to matched human genome regions. Sequin-derived alignments should be down-sampled using the anaquin 'calibrate' tool, which calibrates sequin coverage to corresponding human genome regions. Corresponding human and sequin regions can then be analyzed in parallel using standard bioinformatic tools to evaluate the detection of variants. Coverage calibration ensures a fair comparison between sequin and sample regions.

Materials

Biological materials

! CAUTION The cell lines used in your research should be regularly checked to ensure that they are authentic and that they are not infected with mycoplasma. **▲ CRITICAL** To ensure purity of the sample materials, cell lines used in the WGS and targeted sequencing demonstrations were obtained from the American Type Culture Collection (MCF7, K562) and Coriell Institute for Medical Research (NA12878) and cultured in a dedicated facility, and DNA was extracted using standard methods (DNeasy Blood and Tissue Kit, Qiagen, cat. no. 69504).

- NA12878 cells (Coriell Institute, cat. no. GM12878)
- MCF7 cells (ATCC, cat. no. HTB-22)
- K562 cells (ATCC, cat. no. CCL243)

Reagents

▲ CRITICAL Sequin standards are compatible with any library preparation, target-enrichment strategy or sequencing platform. Users are free to source reagents from alternative vendors that can be easily substituted and used in the following workflow. **▲ CRITICAL** To reduce the risks of experimental cross-contamination, stock reagents used in the preparation of pre-amplification DNA, WGS and targeted sequencing libraries should be maintained independently for each preparation stage.

General

- Nuclease-free water (Thermo Fisher Scientific, cat. nos. AM9937 and AM9932)
- Ethanol (molecular biology grade; Sigma-Aldrich, cat. no. E7023-500ML)
- Hydrochloric acid (Merck, cat. no. 320331) **! CAUTION** Hydrochloric acid, used to adjust the pH of Tris-HCl, is highly corrosive and should be handled inside a fume hood while wearing appropriate personal protective equipment.
- Trizma base (Merck, cat. no. T6791)

Sequins (Genome v2)

▲ CRITICAL We recommend using the Genome v2 Mix for all applications (including WGS and targeted sequencing) apart from paired tumor/normal analysis for somatic variant detection, wherein the Matched Normal v2 Mix should instead be added to the matched germline DNA sample. In this protocol, we use the Genome v2 Mix during single-sample WGS (Step 2A) and both mixtures during targeted sequencing of the mock tumor/normal sample pair (Step 2B).

- Genome v2 Mix (lyophilized DNA; http://www.sequinstandards.com)
- Matched Normal v2 Mix (lyophilized DNA; http://www.sequinstandards.com)

WGS

- TruSeq DNA PCR-free Library Preparation Kit (Illumina, cat. no. FC-121-3003)
- DNA High Sensitivity Reagent Kit (Perkin-Elmer, cat. no. CLS760672)
- HiSeq X Ten Reagent Kit (Illumina, cat. no. FC-501-2501)

Targeted sequencing

- KAPA HyperPlus Kit (KAPA Biosystems, cat. no 07 962 401 001)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63880)
- Agilent High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626)
- SeqCap EZ Accessory Kit v2 (Roche NimbleGen, cat. no. 07 145 594 001)
- SeqCap EZ Developer Enrichment Kit (Roche NimbleGen, cat. no. 06 471 684 001) ▲**CRITICAL** Stock biotinylated probes (SeqCap EZ Developer Enrichment Kit) should be divided into aliquots of single-use 4.5-μl working volumes to avoid freeze–thaw degradation.
- SeqCap EZ Hybridization Kit (Roche NimbleGen, cat. no. 05 634 261 001)
- SeqCap HE-Oligo Kits A and Kit B(Roche NimbleGen, cat. nos. 06 777 287 001 and 06 777 317 001)
- SeqCap Adapter Kits A and Kit B (Roche NimbleGen, cat. nos. 07 141 530 001 and 07 141 548 001)
- Dynabeads M-270 Streptavidin (Invitrogen, cat. no. 65305)

qPCR

- Power SYBR Green PCR Master Mix (Thermo Fisher Scientific, cat. no. 4367659)

Equipment

General

- Single-channel pipettes (2 μl; Gilson, cat. no. F144801)
- Single-channel pipettes (20 μl, 200 μl and 1,000 μl; Gilson, cat. no. F167300)
- Pipette tips (barrier RNase/DNase-free; 10, 30, 200 and 1,000 μl; Gilson, cat. nos. F171203, F171303, F171503 and F171703, respectively)
- PCR tubes (0.2 ml; Eppendorf, cat. no. 0030124332)
- LoBind nuclease-free microfuge tubes (1.5 ml; Eppendorf, cat. no. 0030108051)
- Microcentrifuge (Bio-Rad, cat. no. 1660613)
- Vortex mixer (Bio-Rad, cat. no. 1660621)
- DynaMag-2 magnet (Thermo Fisher Scientific, cat. no. 12321D)
- Precision needle (18 gauge; Becton Dickinson, cat. no. 302032)
- Centrifuge tubes (15 ml; Merck, cat. no. CLS430055)

For WGS

- LE220 focused-ultrasonicator with AFA Technology (e.g., Covaris, cat. no. 500219)
- LabChip GX Touch System (Perkin-Elmer, cat. no. CLS138162)
- Sequencing system (X Ten platform; Illumina, cat. no. SY-412-1001)

Targeted sequencing

- Eppendorf Vacuum Concentrator Plus (Eppendorf, cat. no. 5305000380)
- 2100 Bioanalyzer instrument (Agilent Technologies, cat. no. G2939BA)
- C1000 Touch thermal cycler with dual 48-48 fast-reaction module (Bio-Rad, cat. no. 185-1148)
- Digital dry bath (Bio-Rad, cat. no. 166-0563)
- HiSeq 2500 sequencing system (Illumina, cat. no. SY-401-2501)

qPCR

- ABI PRISM 7900HT sequence detection system (Thermo Fisher Scientific, cat. no. 4317596)
- MicroAmp optical 384-well reaction plate (Thermo Fisher Scientific, cat. no. 4309849)
- Optically clear sealing tape (Sarstedt, cat. no. 95.1994)

Box 3 | Anaquin internal resource files explained

By default, anaquin uses a set of internal resource files that include sequin sequences and gene and variant annotations. These resource files are internally located within the 'resources' directory (that should be located with the anaquin binary under a common parent directory). If required, users can override these default resource files and manually provide their own modified annotations for custom analysis.

The anaquin resource files include:

- sequin_sequences_*.fa: sequences of all sequins contained within the human genome mixture
- sequin_regions_hg38_*.bed: regions of the *hg38* reference genome that are represented by sequins
- sequin_smallvariants_hg38_*.vcf: *hg38* coordinates for SNV and indel variants represented by sequins
- sequin_structural_hg38_*.vcf: *hg38* coordinates for structural variants represented by sequins
- sequin_features_hg38_*.bed: regions of *hg38* represented by sequins and stratified according to a range of features, including GC content and repeats
- sequin_features_hseq_*.bed: regions of sequin sequences stratified according to a range of features, including GC content and repeats
- sequin_hlaregions_hg38_*.bed: coordinates for *HLA* scaffold regions from the *hg38* reference genome assembly that are represented by sequins
- synthetic_ladder_*.tsv: specifies the copy number of sequin elements that constitute a synthetic ladder measuring DNA copy number

Bioinformatic analysis

- High-performance computing resources with a recommended minimum 16 GB of RAM and 8 CPU cores for processing NGS data ▲**CRITICAL** Additional computing resources will reduce the time required to complete some of the bioinformatic steps.
- Ubuntu (v.14.04) operating system or equivalent (https://www.ubuntu.com/download/desktop) ▲**CRITICAL** Although Ubuntu was used for the data analysis steps in this protocol, other Unix/Linux distributions can alternatively be used.

Data files

- Users can access example sequin data files in order to run the bioinformatic steps in this protocol (Steps 12–25). To do so, a user account must first be created at http://www.sequinstandards.com. After registering, users should visit http://www.sequinstandards.com/resources/#nature_protocols and download files to their local working directory as detailed in the 'Equipment setup' section.

Anaquin software

- We have developed a software toolkit, named anaquin[18], to facilitate the analysis of sequins. This includes tools to partition sequin-derived reads from the accompanying sample ('split'), calibrate sequin alignment coverage to match the accompanying sample ('calibrate') and calculate diagnostic statistics regarding variant calling ('germline' and 'somatic'). The software is implemented in C++/R and is freely available under a BSD license. The source code is available from https://github.com/sequinstandards/RAnaquin. Anaquin is compatible with most common bioinformatic tools and accepts standardized file formats (FASTQ, BAM, VCF, BED), allowing easy integration into standard NGS workflows. By default, anaquin utilizes a set of internal resource files (Box 3) that accompany the software package and minimize the user input required to run anaquin tools. However, users can manually provide their own modified reference files for custom analyses.

Other software for NGS analysis ▲**CRITICAL** There are a wide range of software tools available for analysis of NGS data and identification of variants. In this protocol, we use a range of popular third-party software tools (see below). However, users should feel free to use alternative software tools. We recommend that users familiarize themselves with the use of third-party tools, as optional parameters can influence analytical performance. Within the workflow, we have simply used the default parameters for clarity. However, parameter optimization may improve results (indeed, sequins can be used to evaluate and optimize the parameters used).

- BWA (v.0.7.16; http://bio-bwa.sourceforge.net)
- SAMtools (v.1.9; http://www.htslib.org)
- Strelka2 (v.2.9.2; https://github.com/Illumina/strelka)
- GATK (v.4.0.0; https://www.broadinstitute.org/gatk/download)
- Trim Galore (v.0.5.0; http://www.bioinformatics.babraham.ac.uk/projects/trim_galore)
- Integrative Genomics Viewer (IGV v.2.4.7; http://software.broadinstitute.org/software/igv/)
- R (v.3.4.2; https://cran.r-project.org)

- Bioconductor (v.3.5; https://bioconductor.org)
- Python (v.2.7.0; https://www.python.org)

Reagent setup

10 mM Tris-HCl, pH 8

Prepare a 1 M stock solution by dissolving 60.55 g of Trizma base in 400 ml of nuclease-free water and adjusting the pH to 8 with hydrochloric acid (~21 ml). Bring the final volume to 500 ml with nuclease-free water. The 1 M stock solution is stable for a minimum of several months and should be stored at room temperature (RT; 18–24 °C). Make a fresh 10 mM working solution on the day of use by diluting 10 µl of the stock solution in 990 µl of nuclease-free water.

Ethanol (80%)

Ethanol (80% vol/vol) should be made fresh on the day of use by combining 8 ml of 100% ethanol with 2 ml of nuclease-free water in a 15-ml centrifuge tube.

Equipment setup

Equipment location

Maintenance of sample temperature and prompt execution of steps during hybridization and washing are critical to successful target enrichment. We recommend that the required equipment be co-localized to facilitate efficient sample transfer and minimize sample cooling. If multiple capture reactions are being performed, they should be staggered sequentially to limit temperature loss.

Thermocycler programming

Thermocyclers should be programmed with the requisite programs in advance of use; we recommend using a PCR thermocycler with a heated lid (such as the Bio-Rad C1000 Touch Cycler) to prevent evaporation during incubation steps.

Temperature equilibration

We recommend preheating the dry baths and equilibrating the wash buffers to the appropriate temperature in advance to assist in sample processing.

Anaquin installation

Anaquin is a dedicated software package for sequin analysis. We recommend working with the latest version of anaquin. To do so, users should visit https://www.sequinstandards.com/software/ and download anaquin to a convenient local directory. The software can be unpacked and installed with the following commands:

```
$ unzip anaquin_*.zip
$ cd anaquin_*
$ make
```

To test the installation, please run the following command:

```
$ anaquin
```

Installation of the anaquin Bioconductor package

Anaquin produces several automated reports evaluating different aspects of NGS assay performance. To generate data plots contained within these reports, the user's system must support the R statistical computing environment (to install R, visit https://cran.r-project.org), and anaquin's accompanying Bioconductor package must be installed. The anaquin Bioconductor package can be downloaded, along with anaquin, from https://www.sequinstandards.com/software/ and can be installed with the following command:

```
$ R CMD INSTALL RAnaquin_*.tar.gz
$ Rscript -e "library(Anaquin); packageVersion('Anaquin')"
```

Users who prefer to install the Bioconductor package from source code, can alternatively use the following commands:

```
# first open the R console
$ R
# now install and load anaquin
> install.packages("devtools")
> devtools::install_github("sequinstandards/RAnaquin")
> library("Anaquin")
```

▲CRITICAL If the anaquin Bioconductor package is not installed, automated performance reports can still be generated; however, data plots may not be visible.

Python installation

Python is also required for report generation. To check your Python installation, run the following commands:

```
# make sure Python is available
$ python
```

Downloading data files

Download the following files from http://www.sequinstandards.com/resources/#nature_protocols:

- Example FASTQ paired-end sequencing libraries (~40 Gb) generated by WGS analysis of NA12878 gDNA spiked with sequins (Step 2A):
  - NA12878_with_sequins.R1.fq.gz
  - NA12878_with_sequins.R2.fq.gz
- Example FASTQ paired-end sequencing libraries (~12 Gb) generated by targeted sequencing of the mock tumor and normal sample pair spiked with sequins (Step 2B):
  - capture_normal.R1.fq.gz
  - capture_normal.R2.fq.gz
  - capture_tumor.R1.fq.gz
  - capture_tumor.R2.fq.gz
- For the targeted sequencing experiment, users should restrict their analysis to targeted regions of the human genome (and corresponding sequins). Coordinates for the custom gene panel used in this protocol (Step 2B) are defined in the following BED file: capture_panel_regions.hg38.bed ▲CRITICAL This capture panel design file defines coordinates with respect to the hg38 reference genome. If users wish to use an alternative build (e.g., hg19), coordinates can be converted using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).
- To reduce run-time and the risk of potential conflicts, we encourage users to limit their analysis to the primary human chromosomes. To do so, the following BED file (for hg38) and an accompanying index should be downloaded and supplied during variant calling:
  - human_chromosomes.hg38.bed.gz
  - human_chromosomes.hg38.bed.gz.tbi
- We provide a table containing useful information about each individual sequin, such as genotype, allele frequency and the identity of the genetic feature it represents. Although this file is not required to complete this protocol, users are encouraged to access and explore the catalog: sequin_attributes_2.6.tsv
- Users are required to align input FASTQ files to the human reference genome. We encourage users to download the most recent genome build (hg38), available from the GATK resource bundle, via the following command:

```
$ wget https://storage.googleapis.com/genomics-public-data/resources/
broad/hg38/v0/Homo_sapiens_assembly38.fasta
```

- To assist users during the bioinformatic protocol, we provide a number of intermediate data files (FASTQ, BAM and VCF) that make it possible for users to skip time-consuming steps (e.g., sequence alignment). A list of relevant intermediate files available for download can be found in Supplementary Note 1.

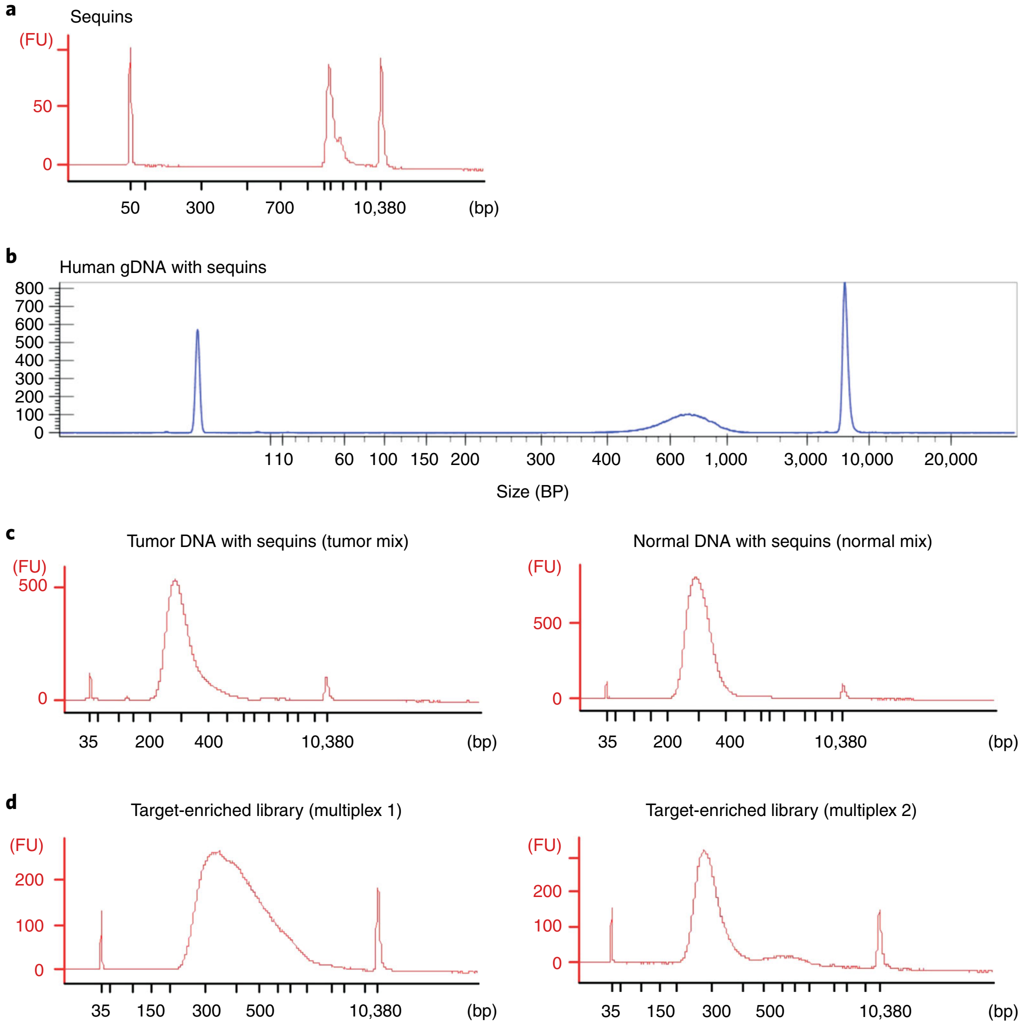


Fig. 6 | Example traces measuring DNA fragment size and abundance. a, Resuspended neat sequin mixture (2100 Bioanalyzer with the 7500 DNA Kit). **b**, Total human genome DNA library containing sequins (LabChip GX Touch System and DNA High Sensitivity Kit). **c**, Total human genome DNA libraries containing sequins (2100 Bioanalyzer with DNA High Sensitivity Kit). Left, patient lung biopsy DNA with genome sequins (tumor); right, patient-matched blood sample with genome sequins (normal). **d**, Captured DNA libraries containing sequins (2100 Bioanalyzer with the DNA High Sensitivity Kit). Observation of appropriate DNA fragment sizes at each stage of the library preparation procedure ensures quality control. FU, fluorescence units.

Procedure

▲**CRITICAL** All laboratory steps (Steps 1–11) should be performed under nuclease-free conditions, with DNA samples and kit reagents additionally thawed and stored on ice before use.

▲**CRITICAL** To reduce the risk of experimental cross-contamination, all pre-PCR amplification steps (Step 2A(i–iii) and 2B(i–ii)) should ideally be performed in a dedicated area or facility, independent of that used for post-PCR amplification stages, including target enrichment (Step 2B(iv–vi)).

Resuspension and storage of sequins ● Timing ~10 min

1 Sequins are DNA molecules that are shipped in a lyophilized format and must be resuspended before use (Fig. 6a). Briefly centrifuge the vials at 2,000*g* for 10 s at RT to collect the dry DNA contents at the bottom of the tube and resuspend the dry sequins in 10 µl of nuclease-free water to generate a stock sequin solution concentration of ~10.0 ng μl^{-1}.

▲**CRITICAL STEP** Failure to centrifuge the tube before resuspension may result in the loss of DNA content.

■ **PAUSE POINT** Following re-suspension, sequin aliquots should be frozen at −20 °C in a frost-free freezer. We would recommend preparing smaller single-use aliquots to minimize the potential impacts of subsequent freeze–thaw cycles. Aliquots can be stored for up to 6 months.

Addition of sequins to samples and library preparation

2 Add sequins to the DNA sample and carry out library preparation using option A for WGS or option B for targeted sequencing.

(A) **Addition of sequins and library preparation for WGS ● Timing ~7 h**

(i) For WGS applications, we recommend adding sequins to your DNA at ~2.0% by mass. In the example below (Step 2A(ii–iv)), add 1 μl of a 2:5 dilution of sequins (Genome v2 Mix) to 200 ng of total human gDNA, i.e., 2.0% by mass.

▲ **CRITICAL STEP** The dilution of sequins in a human DNA sample is proportional to the fraction of reads derived from sequins in the output library. Guidelines on estimating the amount of sequins to be added are provided in the Introduction ('Use of sequins with WGS applications') and Table 1.

(ii) Fragment the combined gDNA sample plus sequins using a sonicator. We use the Covaris LE220, with the following settings: target BP, 450 nt; duty factor, 18%; peak incident power, 450 W; cycles per burst, 200; treatment time, 60 s per column; sample volume, 55 μl; temperature, 5–8 °C.

▲ **CRITICAL STEP** Note that methods such as enzymatic fragmentation (using NEBNext dsDNA Fragmentase M0348 or similar) could alternatively be used to fragment the input DNA.

(iii) Use the fragmented DNA as input for the TruSeq DNA PCR-Free Library Preparation Kit, following the manufacturer's protocol (TruSeq DNA PCR-Free Library Prep Protocol Guide, 15075699Rev.A; https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdnapcrfree/truseq-dna-pcr-free-library-prep-protocol-guide-15075699-a.pdf).

(iv) Verify and quantify the purified libraries on a LabChip GX Touch System using the DNA High Sensitivity Reagent Kit (Fig. 6b). The libraries are now ready for sequencing.

? TROUBLESHOOTING

■ **PAUSE POINT** Libraries can be stored for at least 3 months at −20 °C.

(B) **Addition of sequins and library preparation for targeted sequencing ● Timing ~4 d**

(i) For targeted sequencing applications, we recommend adding sequins to human DNA at ~0.5% by mass (i.e., 1 μl of a 1:20 sequin dilution to 100 ng of total human DNA). Note that the sequin Genome v2 Mix should be added to tumor sample DNA and the sequin Matched Normal v2 Mix to matched normal sample DNA.

▲ **CRITICAL STEP** The addition of sequins to a sample is dependent on the amount of gDNA that is enriched for targeted sequencing. Guidelines on estimating the amount of sequins to be added are provided in the Introduction ('Use of sequins with targeted sequencing applications') and Table 1.

(ii) Carry out enzymatic fragmentation and library preparation using the combined gDNA with sequins as input for the KAPA HyperPlus Kit, in conjunction with the SeqCap Adapter Kits and following the manufacturer's protocol (KAPA Biosystems protocol KR1145, v.4.17; http://netdocs.roche.com/DDM/Effective/0000000000001200000190101_000_03_005_Native.pdf), with seven cycles of PCR amplification at step 5.3 (page 14). For the post-amplification cleanup, use the size-selection steps in section A1 (page 15) in place of section 6 (page 14). At step A1.19 (page 15), use 22 μl of nuclease-free water to liberate the DNA from the Agencourt AMPure XP beads. In addition, at step A1.22 (page 15), carefully transfer 20 μl of the supernatant to a 1.5-ml LoBind microcentrifuge tube.

▲ **CRITICAL STEP** Note that alternative methods, such as sonication, can be used to fragment input DNA before library construction with other preparation kits (e.g., the KAPA Library Preparation Kit, cat. no. 07 137 974 001).

▲ **CRITICAL STEP** If samples will be multiplexed before capture and sequencing, ensure that compatible index combinations (per Roche recommendations) are added during library preparation, while also considering the appropriate corresponding HE Index (blocking) oligos in Step 2B(iv).

▲**CRITICAL STEP** Although the number of cycles used during PCR amplification is largely determined by DNA input amount, the cycle number should be kept to a minimum to avoid generating excessive PCR artifacts.

? TROUBLESHOOTING

(iii) Verify and quantify the purified libraries on an Agilent 2100 Bioanalyzer using the Agilent High Sensitivity DNA Kit (Fig. 6c).

? TROUBLESHOOTING

■**PAUSE POINT** Purified pre-capture libraries can be stored for at least 3 months at −20 °C.

(iv) Use the Roche NimbleGen standard capture protocol for target enrichment (http://netdocs.roche.com/PPM/SeqCapEZLibrarySR_Guide_v3p0_Nov_2011.pdf) with double-capture modification (http://netdocs.roche.com/PPM/Double_Capture_Technical_Note_August_2012.pdf) to ensure higher enrichment of target sequences.

Briefly, combine 1 µg of pre-capture DNA library from Step 2B(iii) with 5 µg of Cot-1 DNA (SeqCap EZ Accessory Kit v2), 1 µl of 1,000 µM Universal oligo and 10 µl total of 100 µM HE Index oligos (SeqCap HE-Oligo Kit A/B).

▲**CRITICAL STEP** If the yield of pre-capture DNA library from Step 2B(iii) is low, a minimum input of 200 ng can be used for capture in Step 2B(iv).

▲**CRITICAL STEP** For samples that are to be multiplexed during capture, consider the ratio at which they are mixed, given that this will determine the fraction of sequenced reads in the final sequenced libraries for each sample. In addition, ensure that each library has a different index and that the appropriate corresponding HE Index (blocking) oligos are used in this step. For example, libraries indexed with SeqCap Index Adapter 2 (A2) will require subsequent blocking with SeqCap HE Index 2 Oligo (B2).

(v) Lyophilize the combined solution at 60 °C in a vacuum concentrator.

(vi) Perform subsequent steps per Roche NimbleGen double-capture protocol step 2.5 (page 8).

▲**CRITICAL STEP** Dynabeads can adhere to plasticware while in the wash buffers. To prevent this, use low vortex speed during washing steps. Beads bound to the walls of the plastic tubes can be scraped and washed back into the solution with a pipette tip.

(vii) Verify and quantify the cleaned captured libraries on an Agilent 2100 Bioanalyzer using the Agilent High Sensitivity DNA Kit (Fig. 6d).

▲**CRITICAL STEP** Unincorporated primers, visible as small peaks close to the lower marker on Agilent Bioanalyzer traces, can adversely impact downstream sequencing and should be removed with a further round of Agencourt AMPure XP bead capture library purification.

■**PAUSE POINT** Purified capture libraries can be stored for at least 3 months at −20 °C.

Validation of library enrichment and composition by qPCR ● Timing ~4 h

3 Upon first use of sequins in an experimental design, users may want to evaluate the success of target enrichment by qPCR to ensure that sequins constitute a minor fraction of the library before sequencing. To do so, first dilute pre-capture (from Step2B(iii)) and post-capture (from Step 2B (vii)) library samples at a ratio of 1:20 with nuclease-free water. Prepare 10 µl of diluted DNA for each target to be assessed.

4 Briefly centrifuge lyophilized target primers (2,000*g*, RT, 10 s) to collect the dry contents at the bottom of the tube and resuspend to make a stock 100 µM solution for each primer in nuclease-free water; store indefinitely at −20 °C. From these, prepare 2 µM working solutions with nuclease-free water and store at 4 °C for up to 1 year.

▲**CRITICAL STEP** In this example, we use validated primer pairs targeting five cancer-associated genes (*ALK, BRAF, PIK3CA, PTEN* and *TP53*) that are typically targeted for enrichment in tumor samples, as well as several human genome regions that are not typically targeted, to provide a semi-quantitative assessment of target enrichment (Table 2). Additional qPCR assessment of an uncaptured housekeeping gene, such as *GAPDH*, is essential to provide intra-sample normalization of subsequent data.

5 For each target, prepare 24.5 µl of qPCR master mix by combining 17.5 µl of Power SYBR Green PCR Master Mix with 3.5 µl of forward primer and 3.5 µl of reverse primer for each sample to be assessed. Prepare a further 24.5 µl of qPCR master mix to generate a no-template control.

Table 2 | qPCR primers for amplification of five commonly enriched human gene targets and their mirrored sequin counterparts

Type	Gene/sequin name	Size (nt)	Forward primer (5′–3′)	Reverse primer (5′–3′)
Human	*PTEN*	104	CATACCAGGACCAGAGGAAACC	CCTTGTCATTATCTGCACGCTC
Human	*BRAF*	82	CCACAAAATGGATCCAGACAACT	AGGTGATTTTGGTCTAGCTACAGT
Human	*TP53*	96	CTCCCCTTTCTTGCGGAGA	AATCTACTGGGACGGAACAGC
Human	*ALK*	82	CAGCAAAGCAGTAGTTGGGG	CTGCAGAGCCCTGAGTACAA
Human	*PIK3CA*	100	TGATGCTTGGCTCTGGAATG	TCCAAAGCCTCTTGCTCAGTT
Sequin	*PTEN*	104	GGAACAGTAATAGACGTGCGAG	GTATGGTCCTGGTCTCCTTTGG
Sequin	*BRAF*	82	TCCACTAAAACCAGATCGATGTCA	GGTGTTTTACCTAGGTCTGTTGA
Sequin	*TP53*	96	TTAGATGACCCTGCCTTGTCG	GAGGGGAAAGAACGCCTCT
Sequin	*ALK*	82	GACGTCTCGGGACTCATGTT	GTCGTTTCGTCATCAACCCC
Sequin	*PIK3CA*	100	AGGTTTCGGAGAACGAGTCAA	ACTACGAACCGAGACCTTAC
Human	*GAPDH*	69	TCGACAGTCAGCCGCATCT	CTAGCCTCCCGGGTTTCTCT

These qPCR primers are used for assessing the enrichment of target genes and the abundance of sequins in libraries before sequencing. GAPDH is not captured on the custom gene panel and provides an off-target control.

6 Pipette 7 µl of qPCR master mix into each of three wells of a MicroAmp optical 384-well reaction plate and dispense into this 3 µl of either the diluted pre-capture or post-capture DNA (or nuclease-free water for the no-template control). This generates a triplicate of a 10-µl qPCR reaction per sample target.

7 Seal the plate with optically clear sealing tape and run the following qPCR cycle on the ABI PRISM 7900HT sequence detection system: 50 °C for 2 min; 95 °C initial denaturation for 10 min; annealing, extension and reading with 40 cycles of 95 °C for 15 s and 60 °C for 70 s.

8 For each sample, deduct the mean cycle threshold (Ct) value for *GAPDH* from those of the target gene(s) to normalize the data (ΔCt) . Log transform the ΔCt value for each target with the following equation: $2^{-\Delta\mathrm{Ct}}$.

9 Calculate the relative abundance of the target gene by dividing post-capture log transformed ΔCt values by those of the matched pre-capture sample (Fig. 7a).

10 Calculate the enrichment of sequin target versus genome target loci by dividing post-capture *GAPDH*-normalized log transformed ΔCt values for sequin targets by those for the corresponding genome targets (Fig. 7b).

▲**CRITICAL STEP** We have mirrored the primers that target cancer-associated genes so they amplify their matched sequin counterparts (Table 2). Given that GC content, annealing temperatures and product size are identical between the endogenous genes and their sequin counterparts, the qPCR results can be directly compared to evaluate the fraction of the library constituting sequins. Enrichment values exceeding those identified in the example dataset should be avoided, as they suggest sequins constitute a large fraction of the captured library.

? TROUBLESHOOTING

Sequencing ● Timing ~6 d

11 Subject the sample to NGS. In this protocol, we have used the Illumina X Ten (for WGS) or HiSeq 2500 (for targeted sequencing) platforms according to the manufacturer's instructions (HiSeq X System Guide, 15050091 v07 (https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseqx/hiseq-x-system-guide-15050091-07.pdf) and HiSeq 2500 System Guide, 15035786 v02 (https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2500/hiseq-2500-system-guide-15035786-02.pdf), respectively).

Trimming of sequencing reads ● Timing ~6 h

12 Users should now have obtained NGS library files in FASTQ format and are ready to begin the bioinformatic workflow. Before proceeding, we recommend that users remove contaminating

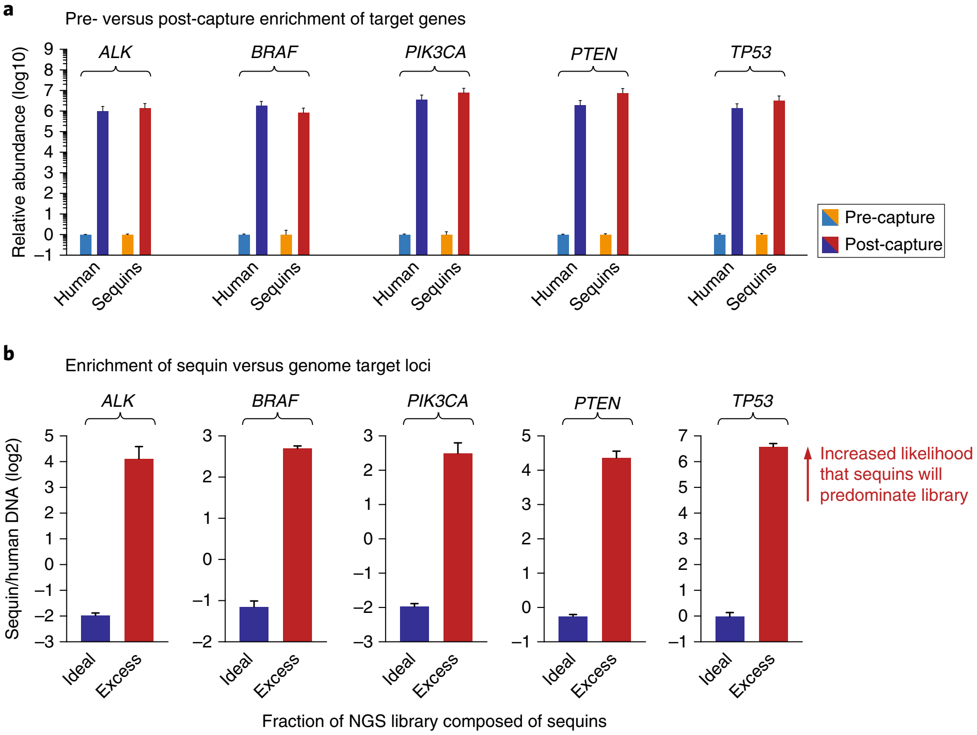


Fig. 7 | Example qPCR assessment of target enrichment for *ALK*, *BRAF*, *PIK3CA*, *PTEN* and *TP53*. a, Comparison of the abundance of endogenous human genes and sequins in libraries before (orange/blue bars) and after target enrichment (purple/red bars). These qPCR analyses confirm successful enrichment of human genes and corresponding sequin targets by capture hybridization. **b**, Abundance of sequins relative to endogenous human genes for recommended sequin spike-in value (0.5%; purple bars) and for samples with excess sequins spiked in (2.5%; red bars). All data are normalized to *GAPDH*. Error bars represent standard error. These qPCR analyses identify samples to which sequins have been added in excess, risking saturation of the resulting sequencing libraries by sequin reads.

adaptor sequences using the popular tool Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) by running the following command:

```
$ trim_galore --paired NA12878_with_sequins.R1.fq.gz NA12878_with_
sequins.R2.fq.gz
```

▲ **CRITICAL STEP** Contaminating adaptor sequences can reduce the specificity of the anaquin 'split' tool, which classifies sequencing reads as sample- or sequin-derived (Step 13), and can interfere with downstream analyses.

Separation of sequin and sample reads ● Timing ~2 h

13 Partition reads derived from (i) sequin standards and (ii) accompanying human sample DNA into separate libraries using the anaquin 'split' tool. 'Split' searches each read-pair for *k*-mers that indicate whether it derives from sequin or sample DNA. In addition, 'split' will also reverse the sequence orientation of sequin-derived reads while preserving Phred quality scores and read-pair relationships, thereby allowing sequin-derived reads to be subsequently aligned to the human reference genome. Optionally, 'split' can also provide a preliminary report containing various quality metrics, based on an alignment-free analysis of sequins within the library. To partition the trimmed WGS library, run the following command:

```
$ anaquin split -t 8 -o split_out --report /
-1 NA12878_with_sequins.R1_val_1.fq.gz -2 NA12878_with_sequins.R2_
val_2.fq.gz
```

In this example command, `-t` enables multithreading (8 CPUs) to speed up the partition process and `-1`/`-2` specify the input FASTQ files. Because the `--report` option was invoked,

Box 4 'Split' | output libraries explained

The Anaquin 'split' tool separates sequencing reads derived from sequin standards and accompanying sample DNA, and reverses the sequence orientation of sequin-derived reads. Sequin reads are further subdivided according to the synthetic genome features that they represent, with separate FASTQ output files created for each sequin category (Box 2). The 'split' output directory contains the following partitioned libraries.

Sample reads (these reads are not modified by 'split')

- split_sample_*.fq.gz: sequencing reads that were derived from sample DNA

Genome sequin reads (these reads are reversed to match the orientation of the human genome)

- split_sequin_*.fq.gz: sequins representing common and disease-associated small variants (including germline and somatic variants), as well as clinically relevant and analytically challenging genome regions
- split_sv_*.fq.gz: sequins representing large structural variants. This includes translocations, deletions, tandem duplications, inversions, viral insertions and mobile elements
- split_immune_*.fq.gz: sequins representing immunoglobulin and T-cell receptors
- split_hla_*.fq.gz: sequins representing human leukocyte antigen alleles
- split_mito_*.fq.gz: sequins representing the human mitochondrial genome

Synthetic sequin reads (these sequin reads do not correspond to human sequences and are not reversed)

- split_ladder_*.fq.gz: an artificial sequence ladder used for measuring DNA copy number
- split_info_*.fq.gz: a unique sequence barcode specifying the identity of the sequin mixture used
- split_vector_*.fq.gz: contaminating reads from vector sequences used in sequin production

'split' will also produce a brief report describing sequin performance (see Supplementary Note 2 to access an example report). 'Split' generates separate libraries containing sample-derived (split_sample.*.fq) and sequin-derived (split_sequin.*.fq) reads in the output directory (split_out/).

Note that the 'split' tool automatically partitions sequin-derived reads into multiple FASTQ libraries that represent the different genomic features (e.g., structural variants, HLA alleles, immune receptors) described in Box 2. In this protocol, we work with only the split_sequin.*.fq.gz library, which includes sequins representing common and disease-associated small variants and clinical genome regions. However, we encourage users to consider other categories for further analysis (Box 4).

? TROUBLESHOOTING

Alignment to human reference genome ● Timing ~8 h

14 After partitioning the WGS library into sample- and sequin-derived reads, both groups should be aligned to the human reference genome (hg38). First, build a set of genome indices (these can be used for all subsequent alignment, variant calling, manipulation and visualization processes) using the following commands:

```
$ bwa index Homo_sapiens_assembly38.fasta
$ samtools faidx Homo_sapiens_assembly38.fasta
$ gatk CreateSequenceDictionary -R Homo_sapiens_assembly38.fasta
```

15 Next, align the sample- and sequin-derived libraries separately to the reference genome using BWA[15]:

```
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
split_out/split_sample_1.fq.gz split_out/split_sample_2.fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > sample.bam
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
split_out/split_sequin_1.fq.gz split_out/split_sequin_2.fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > sequin.bam
```

▲ **CRITICAL STEP** BWA outputs alignments to the console in SAM format and we must therefore redirect the alignments into SAMtools[16] for compression and sorting.

Calibration of sequin coverage to sample coverage ● Timing ~1 h

16 Sequencing coverage is an important variable that impacts many aspects of downstream analysis[2]. Accordingly, the alignment coverage of each sequin should be calibrated to match the alignment

coverage of the counterpart region in the human genome. To do so, we use the anaquin 'calibrate' tool. For each individual sequin region, 'calibrate' measures sample-alignment coverage and sequin-alignment coverage, and then calculates a normalization factor by which to downsample sequin alignments within that region, thereby matching sequin coverage to the human sample. 'Calibrate' will also remove alignments at the terminal nucleotides of each sequin that can be artifactually enriched during adaptor ligation in some library preparation methods (Supplementary Fig. 1). To calibrate sequin alignments to the accompanying human sample, use the following command:

```
$ anaquin calibrate -t 8 -o calibrate_out --sequin sequin.bam --sample
sample.bam
$ mv ./calibrate_out/calibrate_sequin_calibrated.bam ./sequin.cali-
brated.bam
```

Sequencing edge effects cause reductions in coverage at the terminal regions of sequins[23]. By default, 'calibrate' omits terminal regions (~550 nt) from analysis to mitigate this effect. This size of the terminal regions can be adjusted using the `--edge` option. See Supplementary Fig. 1 for an example.

? TROUBLESHOOTING

Identification of germline variants ● Timing ~2 h

17 Next use Strelka2[17] to identify germline variants within both sample- and sequin-derived alignments. As we only briefly outline the required commands here, users are encouraged to consult the Strelka2 documentation for further details (https://github.com/Illumina/strelka/). Use the following commands to generate indices for sample and calibrated sequin BAM alignment files:

```
$ samtools index sample.bam
$ samtools index sequin.calibrated.bam
```

18 Then supply sample and calibrated sequin BAM files separately to Strelka2 as follows, in order to perform germline variant discovery for each library:

```
$ configureStrelkaGermlineWorkflow.py --runDir sample_germline /
--callRegions human_chromosomes.hg38.bed.gz /
--referenceFasta Homo_sapiens_assembly38.fasta --bam sample.bam
$ sample_germline/runWorkflow.py -m local -j 8
$ configureStrelkaGermlineWorkflow.py --runDir sequin_germline /
--callRegions human_chromosomes.hg38.bed.gz /
--referenceFasta  Homo_sapiens_assembly38.fasta  --bam  sequin.cali-
brated.bam
$ sequin_germline/runWorkflow.py -m local -j 8
```

Note that in this example, we limit our analysis to the primary human chromosomes, defined in the file human_chromosomes.hg38.bed.gz, in order to reduce run-time and risk of potential conflicts.

19 Finally, decompress and rename the output files for convenience:

```
$ zcat sample_germline/results/variants/variants.vcf.gz > sample.var-
iants.vcf
$ zcat sequin_germline/results/variants/variants.vcf.gz > sequin.var-
iants.vcf
```

These example commands generate two VCF files, sequin.variants.vcf and sample.variants.vcf, which contain candidate germline variants identified within sequin alignments and the accompanying human sample alignments, respectively. We encourage users to visualize variant candidates and sample/sequin-derived sequencing alignments in a compatible genome browser of their choice (Box 5; Fig. 8).

Box 5 | Visualizing sequins and corresponding human genome regions

Users are encouraged to inspect sequins, particularly in comparison to their counterpart human genome regions, to help interpret variant candidates and identify analytic blind spots. We recommend that users download and install the IGV (available from http://www.broadinstitute.org/igv), which is a popular tool that allows users to inspect sequencing alignments and variant candidates[37].

To view the alignments generated above (Steps 15 and 16): open IGV; select 'File' from the overhead menu and 'Load from File…' from the dropdown options; then locate sample/sequin alignments in the working directory (i.e., sample.bam, sequin.calibrated.bam). Please note that IGV requires a BAM Index (BAI) file for each alignment file, which must be located in the same directory (Step 17).

Variants can also be loaded for viewing using the same process as above, this time locating the VCF files containing sample/sequin variant calls in the working directory (i.e., sequins.variants.vcf, sample.variants.vcf; Steps 18 and 19). Sequin alignments and variant calls can then be inspected and compared to corresponding regions of the human genome to reveal subtle features, such as alignment artifacts. For example, Fig. 8 shows sample and sequin alignments at a human variant that is directly matched by a synthetic sequin variant.

Evaluation of germline variant detection performance ● Timing ~15 min

20 It is now possible to evaluate the sensitivity and accuracy of variant detection with sequins. To do so, use the anaquin 'germline' tool, which classifies germline variant candidates within sequins as true positives (TPs) or false positives (FPs), and identifies known variants that were missed (false negatives; FNs). To launch 'germline', run the following command:

```
$ anaquin germline -o germline_out --report /
--sample sample.variants.vcf --sequin sequin.variants.vcf
```

The 'germline' tool prints several files to the output directory (germline_out), including the following:

- germline_variants.tsv: detailed information about each candidate variant, including a classification label (TP, FP, FN or sample variant candidate (SV)).
- germline_report.html: detailed report on sequin and sample variant detection performance (see Box 6 for details and Supplementary Note 2 to access an example report). This report can be loaded into a standard web browser (e.g., Chrome, Safari).

21 Various criteria can be used to filter variant call-sets, theoretically excluding erroneous calls. By filtering a set of variant candidates and then re-evaluating them with 'germline', it is easy to assess the impact of different filtering strategies on diagnostic performance. For example, use the following commands to exclude variant candidates below an arbitrary confidence threshold (QUAL = 100), before re-evaluating performance:

```
$ gatk SelectVariants -R Homo_sapiens_assembly38.fasta /
-V sample.variants.vcf -O sample.filtered.vcf -select 'QUAL > 100'
$ gatk SelectVariants -R Homo_sapiens_assembly38.fasta /
-V sequin.variants.vcf -O sequin.filtered.vcf -select 'QUAL > 100'
$ anaquin germline -o filtered_out --report --sequin sequin.filtered.
vcf /
--sample sample.filtered.vcf
```

Assessment of somatic variant detection in targeted DNA sequencing ● Timing ~2 h

22 Somatic mutations can exhibit a wide range of variant allele frequencies (VAFs) due to copy-number variation, tumor heterogeneity and sample impurity. Sequins provide a somatic VAF reference ladder that allows users to assess the sensitivity, precision and quantitative accuracy with which variants at different allele frequencies are detected in an NGS library[10] (Box 1). In this protocol, somatic mutations are detected by comparing a mock tumor sample to a matched normal counterpart. Similarly, somatic mutations are detected within sequins by comparing the sequin 'tumor' mixture to the sequin 'normal' mixture, with these having been added to the tumor and normal samples, respectively (Step 2B).

Use the following commands to prepare processed, calibrated BAM alignment files from the tumor/normal sequencing libraries prepared in Step 2B, which can then be used to identify somatic

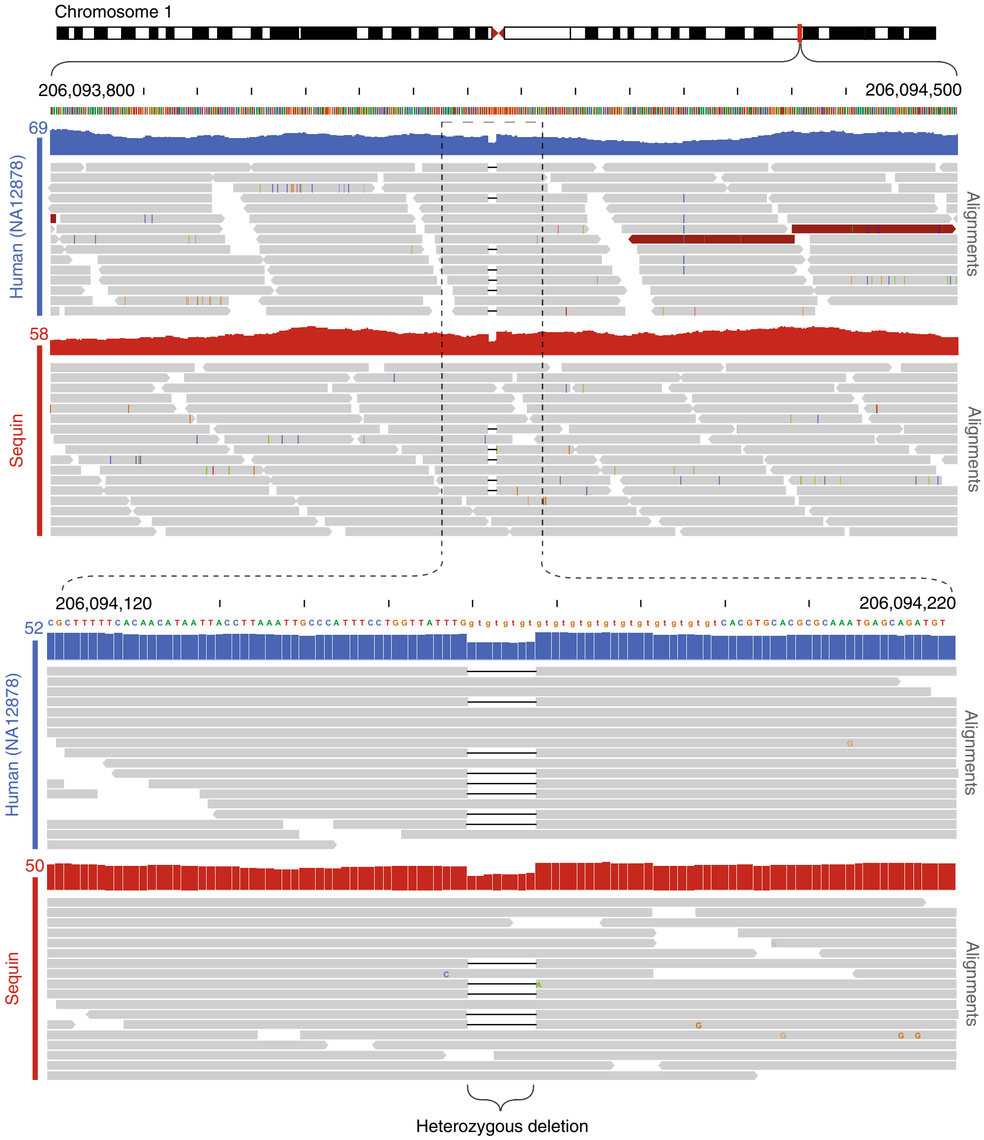


Fig. 8 | Example of sequin and corresponding human variants. Genome browser view (from IGV) of sequin alignments (red) and human alignments (blue) at a heterozygous deletion from the NA12878 genome that is directly represented by a corresponding sequin standard. Sequin alignments were sub-sampled to achieve coverage matched to the human genome, as shown by the coverage scale bars. Alignment profiles and detection sensitivity/precision are similar for sequin and sample variants, with synthetic sequin variants providing matched examples to assist in interpretation of candidate variants in the accompanying sample. See Box 5 for a discussion of visualization of sequins and corresponding human genome regions.

mutations. Alternatively, pre-processed BAM files for tumor (tumor.sample.bam, tumor.sequin.calibrated.bam) and normal (normal.sample.bam, normal.sequin.calibrated.bam) samples can be downloaded directly (Supplementary Note 1).

```
# trim reads to remove adaptor contamination
$ trim_galore --paired capture_tumor.R1.fq.gz capture_tumor.R2.fq.gz
$ trim_galore --paired capture_normal.R1.fq.gz capture_normal.R2.fq.
gz
# partition sample- and sequin-derived reads
$ anaquin split -t 8 -o tumor.split_out --report /
-1 capture_tumor.R1_val_1.fq.gz -2 capture_tumor.R2_val_2.fq.gz
```

Box 6 | Automated performance reporting

The `--report` option can be invoked when using several anaquin tools ('split', 'germline' and 'somatic'), in order to generate a report (*_report.html) describing various aspects of sequin performance. This report can be viewed in any standard browser (e.g., Chrome, Safari). To do so, open the browser and select 'File' from the overhead menu and 'Open' from the dropdown options; then locate the file *_report.html in the output directory of your desired anaquin tool. The report generated by each anaquin tool will contain different information, which is specific to the tool used and/or the input files supplied. Example reports generated during this protocol are available for download (Supplementary Note 2).

'Split' can generate a preliminary report (split_report.html; Step 13) when it performs a *k*-mer-based partition of sample- and sequin-derived reads. This report is generated directly from the user's FASTQ library and describes the raw sequencing library before alignment, calibration, variant calling or other processes. Given the ease and immediacy with which this report can be generated, it can be used to routinely monitor performance of laboratory steps in the NGS workflow and to enable rapid troubleshooting and quality control of experimental variables.

'Germline' will optionally generate a report that evaluates the detection of germline variants in sequins and the accompanying human sample (germline_report.html; Step 20). Performance metrics such as sensitivity, precision and false-positive rate (per kb) are calculated based on VCF files supplied by the user. In addition to these global metrics, the 'germline' report stratifies variants according to type (e.g., SNV versus indel) and/or genomic context (e.g., local GC content) to provide more nuanced performance statistics.

'Somatic' will optionally generate a report that evaluates the detection of somatic mutations in sequins and the accompanying human sample (somatic_report.html; Step 25). In addition to global metrics calculated based on the user supplied sequin/sample somatic mutation VCF files, the 'somatic' report evaluates the quantitative accuracy, sensitivity and specificity with which mutations at different allele frequencies are detected.

Once open, an anaquin report contains multiple sections compressed into dropdown menus. The 'Anaquin—log reports' section, which appears first in each report, includes names and locations for the files used to generate the report (including internal anaquin files), as well as other useful information pertaining to the analysis.

```
$ anaquin split -t 8 -o normal.split_out /
-1 capture_normal.R1_val_1.fq.gz -2 capture_normal.R2_val_2.fq.gz
# align sample-derived reads to the human reference genome
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
tumor.split_out/split_sample_1.fq.gz tumor.split_out/split_sample_2.
fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > tumor.sample.bam
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
normal.split_out/split_sample_1.fq.gz   normal.split_out/split_sam-
ple_2.fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > normal.sample.bam
# index sample-derived alignments
$ samtools index tumor.sample.bam
$ samtools index normal.sample.bam
# align sequin-derived reads to the human reference genome
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
tumor.split_out/split_sequin_1.fq.gz      tumor.split_out/split_se-
quin_2.fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > tumor.sequin.bam
$ bwa mem -t 8 Homo_sapiens_assembly38.fasta /
normal.split_out/split_sequin_1.fq.gz    normal.split_out/split_se-
quin_2.fq.gz | /
samtools view -@ 8 -b | samtools sort -@ 8 > normal.sequin.bam
# calibrate sequin alignments to match sample coverage
$ anaquin calibrate -o tumor.calibrate_out /
--restrict_regions capture_panel_regions.hg38.bed /
--sequin tumor.sequin.bam --sample tumor.sample.bam
$ mv ./tumor.calibrate_out/calibrate_sequin_calibrated.bam ./tumor.
sequin.calibrated.bam
$ anaquin calibrate -o normal.calibrate_out /
--restrict_regions capture_panel_regions.hg38.bed /
--sequin normal.sequin.bam --sample normal.sample.bam
```

```
$ mv ./normal.calibrate_out/calibrate_sequin_calibrated.bam ./normal.sequin.calibrated.bam
# index calibrated sequin-derived alignments
$ samtools index tumor.sequin.calibrated.bam
$ samtools index normal.sequin.calibrated.bam
```

▲ **CRITICAL STEP** The sequin Matched Normal v2 Mix provides unmutated background sequence to enable paired tumor/normal somatic variant calling; however, users who do not wish to perform paired variant calling can simply call somatic variants in the sequin Genome v2 Mix (and tumor DNA sample) alone.

23 Use Strelka2[17] to identify somatic mutations in sample- and sequin-derived alignments, as follows:

```
$ configureStrelkaSomaticWorkflow.py --exome --ref Homo_sapiens_assembly38.fasta /
--normalBam normal.sample.bam --tumorBam tumor.sample.bam /
--runDir somatic_sample
$ somatic_sample/runWorkflow.py -m local -j 8
$ configureStrelkaSomaticWorkflow.py --exome --ref Homo_sapiens_assembly38.fasta /
--normalBam normal.sequin.calibrated.bam /
--tumorBam tumor.sequin.calibrated.bam /
--runDir somatic_sequin
$ somatic_sequin/runWorkflow.py -m local -j 8
```

24 Strelka produces separate output files containing SNV and indel variant calls; combine these into a single file using the GATK tool 'MergeVcfs' as follows:

```
$ gatk MergeVcfs -O sequin.somatic_variants.vcf /
-I somatic_sequin/results/variants/somatic.snvs.vcf.gz /
-I somatic_sequin/results/variants/somatic.indels.vcf.gz
$ gatk MergeVcfs -O sample.somatic_variants.vcf /
-I somatic_sample/results/variants/somatic.snvs.vcf.gz /
-I somatic_sample/results/variants/somatic.indels.vcf.gz
```

25 To evaluate the detection of somatic mutations in sequins and the accompanying tumor sample, use the anaquin tool 'somatic', which classifies somatic mutation candidates within sequins as TPs or FPs, and identifies FNs. To launch 'somatic', run the following command:

```
$ anaquin somatic --restrict_regions capture_panel_regions.hg38.bed /
-o somatic_out --report --sequin sequin.somatic_variants.vcf /
--sample sample.somatic_variants.vcf
```

In this example, we have restricted our analysis to genome regions that were represented on our target-enrichment panel (`capture_panel_regions.hg38.bed`) by invoking the `--restrict_regions` option. This is useful for targeted sequencing experiments, in which only a fraction of the genome is analyzed, and prevents any confounding effects from off-target alignments. The `--restrict_regions` option could be used to further restrict the analysis according to the user's preferences, for instance, to protein-coding exons, by providing a custom BED annotation.

The 'somatic' tool outputs several useful files that describe the performance of sequins. This includes a detailed report (report.html) that compares the detection of sequin and sample variants in corresponding regions of the genome, providing a ready assessment of diagnostic performance (see Box 6 for details and Supplementary Note 2 to access an example report). Additional output files include summary statistics describing overall performance (somatic_summary.stats) and detailed statistics for each synthetic mutation (somatic_variants.tsv).

Troubleshooting

Troubleshooting advice can be found in Table 3.

Table 3 | Troubleshooting table

Step	Problem	Possible reason	Solution
2A(iv) or 2B(iii)	WGS or pre-capture libraries are of lower than expected yield	DNA input amount is too low	Increase input DNA amount or increase the number of PCR amplification cycles
		DNA is poor quality	Increase the number of PCR amplification cycles, or assess the quality of input DNA using Bioanalyzer or similar
2B(ii)	DNA is not fragmenting to the appropriate size, or not is not fragmenting at all	DNA input used in the fragmentation steps is more—or less—than the 100 ng used in the example	Use the requisite 100 ng of input DNA or re-optimize the fragmentation time
		Alternatively, EDTA is present in the buffer of the source DNA and is inhibiting the action of the fragmenting enzyme	Re-purify the DNA and elute in nuclease-free water
3–10	Low differential qPCR values following target enrichment	qPCR primers used are not specific to captured sequins/genome regions	Design qPCR primers appropriate to captured sequins/genome regions
		Failure to maintain hybridization and washing temperatures	Ensure rapid and efficient sample transfer to avoid temperature loss, or ensure that the equipment used is not faulty
		Degradation of reagents used	Ensure reagents are within use-by date and avoid excessive freeze–thaw cycles
13	A high fraction of the NGS library reads correspond to sequins	Too many sequins added to the sample	Reduce the sequin amount added to the sample and use qPCR validation to assess the library fraction constituting sequins
	Few sequenced reads in the NGS library correspond to sequins	Too few sequins added to the sample	Increase the sequin amount added to the sample and use qPCR validation to assess the library fraction constituting sequins
16	Artifactual alignments are enriched at the termini of sequins	Library preparation causes preferential ligation of adaptors to sequin termini	Apply anaquin ‘calibrate’ to omit alignments that occur at sequin termini

Timing

Step 1, resuspension and storage of sequins: ~10 m
Step 2A(i–iii), WGS library preparation: ~6 h
Step 2A(iv), WGS library validation: ~1 h
Step 2B(i–ii), pre-capture library preparation: ~7 h
Step 2B(iii), pre-capture library validation: ~1 h
Step 2B(iv–vi), hybridization and library capture: 3 d (day 1, ~2 h; day 2, ~5.5 h; day 3, ~3.5 h)
Step 2B(vii), capture library validation: ~1 h
Steps 3–10, qPCR validation: ~4 h
Step 11, sequencing on the Illumina HiSeq 2500 or X Ten: ~6 d
Step 12, trimming of sequencing reads: ~6 h
Step 13, separation of sequin and sample reads: ~2 h
Steps 14 and 15, alignment to human reference genome: ~8 h
Step 16, calibration of sequin alignment coverage to sample: ~1 h
Steps 17–19, identification of germline variants: ~2 h
Steps 20 and 21, evaluation of germline variant detection performance: ~15 m
Steps 22–25, assessment of somatic variant detection in targeted DNA sequencing data: ~2 h

Anticipated results

Here, we describe the results anticipated from the example datasets and bioinformatic workflows used in this protocol. Actual results generated may vary if users implement different versions of the recommended software or alternative genome assemblies, or use their own input datasets. Output

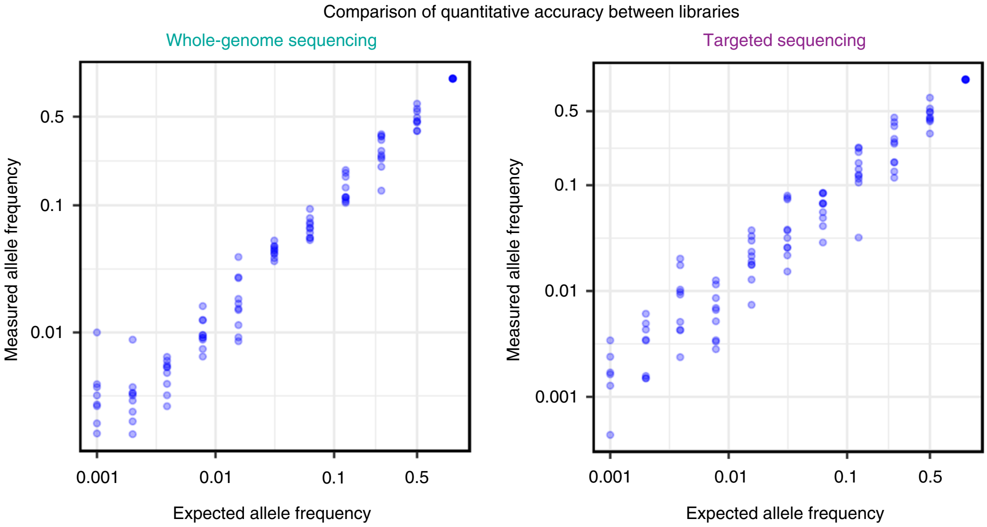


Fig. 9 | Alignment-free comparison of quantitative accuracy between libraries. Scatter plots showing the observed relative to expected variant allele frequency as generated by the anaquin 'split' report function, for both the WGS (left) and targeted sequencing (right) assays described in this protocol. This ladder indicates the quantitative accuracy and sensitivity of the sequenced library before alignment and variant detection. Notably, targeted sequencing library exhibits greater variation and lower accuracy in the measurement of variant frequencies as compared with the WGS library. WGS, whole-genome sequencing.

files generated during the workflow described in this protocol are available for download (Supplementary Note 1).

Alignment-free performance metrics

This protocol describes the analysis of human DNA samples with sequins added, using either WGS or targeted sequencing. In each case, sample- and sequin-derived sequencing reads were initially partitioned using anaquin 'split' (Steps 13 and 22). During this process, 'split' optionally creates a preliminary report that is generated from a direct analysis of the input FASTQ files and provides simple performance and quality-control metrics for the library under analysis (Box 6). See Supplementary Note 2 to access an example report.

The 'split' report includes an alignment-free evaluation of variant sensitivity and quantitative accuracy, providing a simple indication of assay performance that is not biased by the choice of downstream tools for sequence alignment and variant detection, but simply reflects the composition of the sequenced library. This allows the user to quickly evaluate the impact of technical variables during library preparation and sequencing, or to identify defective libraries.

To illustrate this application, in Fig. 9, we present a side-by-side comparison of the VAF ladder generated by 'split' from the WGS and targeted sequencing libraries in this protocol. As expected, we observe increased variability in VAF quantification that results from the capture enrichment process during the targeted sequencing protocol, in comparison to the WGS protocol.

Germline variants identified with WGS

After performing WGS and partitioning sample- and sequin-derived reads, Strelka2 was used to identify germline variants (SNVs and indels) in sequin standards and the accompanying human sample (NA12878; Steps 17–19). Within human genome regions that are represented by sequins ($n = 549$; ~652 kb after excluding terminal regions), 500 candidate variants were detected in sequin alignments and 1,143 in sample alignments.

The anaquin 'germline' tool classifies candidate variants detected in sequin standards as either TPs or FPs, and calculates diagnostic statistics accordingly (Steps 20 and 21). With the workflow described above, 417/422 TP and 54 FP SNVs were detected, yielding sensitivity (sn) and precision (pr) scores of 0.99 and 0.88, respectively. The performance for indels was relatively poor by comparison, with 28/100 TPs and 1 FPs detected (sn = 0.28, pr = 0.97).

Box 7 | Stratifying results by genomic context

Variant detection with NGS is impacted by variant type and surrounding sequence context. For example, it is typically more difficult to identify indels than SNVs, and can be challenging to detect variants that occur in GC-rich or repetitive sequences. Accordingly, diagnostic performance is typically lower for these difficult variants or regions.

Anaquin allows users to easily stratify their results according to such contextual attributes, reporting diagnostic performance for particular variant types or regions. To do this, users are required to provide a BED annotation that stratifies the human genome sequence into regions of interest. Anaquin will then separately report diagnostic performance statistics, such as TP and FP rates, within the supplied regions.

By default, anaquin intersects variant candidates with feature annotations defined in the internal resource file sequin_features_hg38_*.bed. This file stratifies human genome regions represented by sequins into different 'feature categories' ('GCcontent', 'GeneContext', 'MobileElement' and 'SimpleRepeat'), each of which is further subdivided into 'feature types' (e.g., GCrich/ATrich).

However, users can instead provide their own custom annotation containing other relevant features. This can be done by modifying the sequin_features_hg38_*.bed file within the resource distribution. Custom features should be provided in four-column BED format, using the hierarchical naming system: `FeatureCategory_FeatureType`. For instance, annotations for gene-regulatory elements could be added using the following naming system: `RegulatoryElement_Promoter / RegulatoryElement_Enhancer`. Note that although different feature categories may overlap, different feature types within a category must not (Fig. 10).

To illustrate the effect of variant filtering, diagnostic performance was then re-evaluated following the exclusion of variants below an arbitrary confidence threshold (QUAL <100). This filtering strategy excluded all FP sequin SNVs and indels (pr = 1.00) but also results in reduced sensitivity for both categories (sn = 0.92, sn = 0.22). By applying the same filtering strategy to human variant candidates, 91 low-confidence candidates were excluded, retaining a total of 852 SNVs and 200 indels. This simple scenario demonstrates how sequins can be used to inform and optimize variant filtering strategies.

The 'germline' tool further parses sequin and sample variants into subgroups associated with specific sequence features (Box 7; Fig. 10), allowing the impact of variant context on diagnostic performance to be assessed. For instance, only 24/71 synthetic variants occurring at simple repeat sites were detected (sn = 0.34), compared to 421/451 non-repetitive variants (sn = 0.93), highlighting the difficulty of diagnosing repeat-associated variants with NGS.

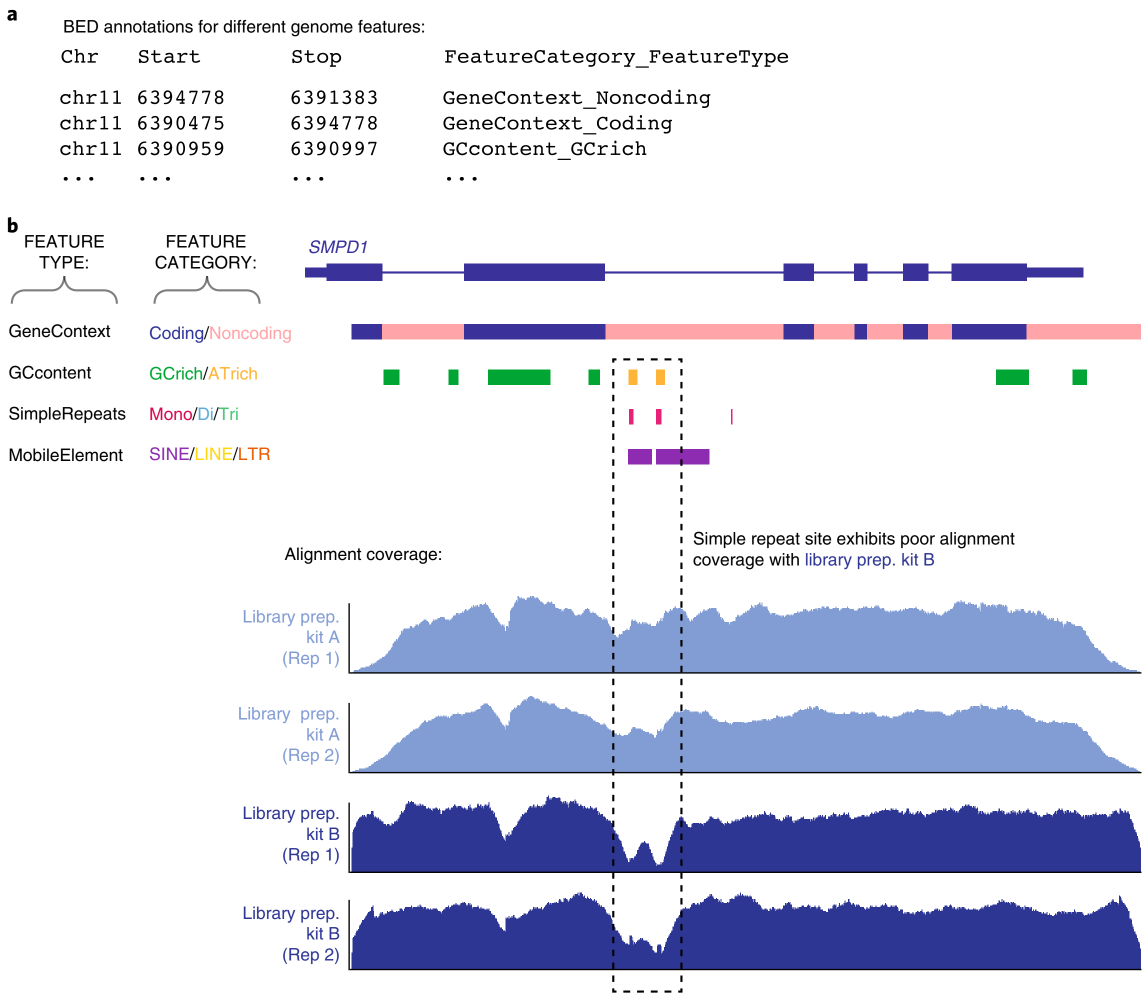
In a typical experiment, these performance statistics are determined only for sequin variants, as the validity of any given variant in the accompanying sample is unknown. However, within this protocol, we analyzed the well-characterized human NA12878 sample[13]. Therefore, we can also assess performance by comparison to high-confidence variants annotated by Genome in a Bottle (GIAB)[24]. Notably, there are 395 sequin variants that contain a matched counterpart in the NA12878 genome, allowing a direct comparison of sequin and NA12878 variants.

Of the 367 matched variants, 332 were detected in sequin-derived alignments, yielding sn scores of 0.90 and 0.30 for SNVs and indels, respectively. By comparison, 326 of the matched variants were detected in the accompanying NA12878 sample, yielding sn scores of 0.88 and 0.30 for SNVs and indels, respectively (Table 4). Within the region of comparison (the intersection of the GIAB high-confidence regions and relevant sequin regions; ~166 kb), five FPs were detected in sequin-derived alignment, as compared to 13 unannotated variants in the accompanying human sample (which are likely to be FPs). This analysis confirms that sequin variants perform similarly to matched germline variants in the accompanying sample in a standard WGS workflow.

More detailed summary statistics, including information on variant types and contexts, can be found in the germline_report.html and germline_summary_stats.tsv files that are generated within the germline_out/ and filtered_out/ directories (Steps 20 and 21). See Supplementary Note 2 to access an example report.

Somatic variants identified with targeted sequencing

This protocol describes the analysis of DNA from a mock tumor sample comprising a mixture of NA12878, MCF7 and K562 cell lines[14], as well as a matched normal sample (NA12878 only) by targeted sequencing. After partitioning both samples into sample- and sequin-derived reads, Strelka2 was used to identify somatic variants in sequin standards and the accompanying tumor sample (Steps 23 and 24). Within captured human genome regions that are directly represented by sequins (~113 kb), 243 somatic variant candidates were detected in sequin alignments and 375 in human alignments.



BED annotations for different genome features:

```
Chr    Start      Stop       FeatureCategory_FeatureType

chr11  6394778    6391383    GeneContext_Noncoding
chr11  6390475    6394778    GeneContext_Coding
chr11  6390959    6390997    GCcontent_GCrich
...    ...        ...        ...
```

Fig. 10 | Impact of sequence context on NGS performance. a, Example lines from the anaquin resource file sequin_features_hg38_2.5.bed, which is used to stratify human genome regions represented by sequins into different feature categories (e.g., GeneContext) and feature types (e.g., coding/noncoding). **b**, Comparison of sequencing coverage within a single sequin (~3 kb) representing the human gene *SMPD1* and analyzed with four different library preparation protocols (not described here). The impact of specific sequence features, annotated above, on sequencing coverage can easily be assessed for each preparation. See Box 7 for a discussion of Stratifying results by genomic context.

Table 4 | Comparison of germline variant detection in sequins and accompanying NA12878 sample

	Combined		**SNVs**		**Indels**	
	Sequins	**NA12878**	**Sequins**	**NA12878**	**Sequins**	**NA12878**
Matched variants	367	367	317	317	50	50
Detected	332	326	317	311	15	15
False positives	5	13	5	13	0	0
Sensitivity	0.90	0.89	1.00	0.98	0.30	0.30
95% CI	0.87–0.93	0.85–0.92	0.98–1.00	0.96–0.99	0.18–0.47	0.18–0.47
Precision	0.99	0.96	0.98	0.96	1.00	1.00
95% CI	0.97–0.99	0.94–0.98	0.96–0.99	0.94–0.98	1.00–1.00	1.00–1.00

95% confidence intervals (CIs) for sensitivity are exact Clopper–Pearson intervals, and for precision are standard logit intervals. Note that the analysis is limited to the intersection of GIAB high-confidence regions and sequin comparison regions (~177 kb).

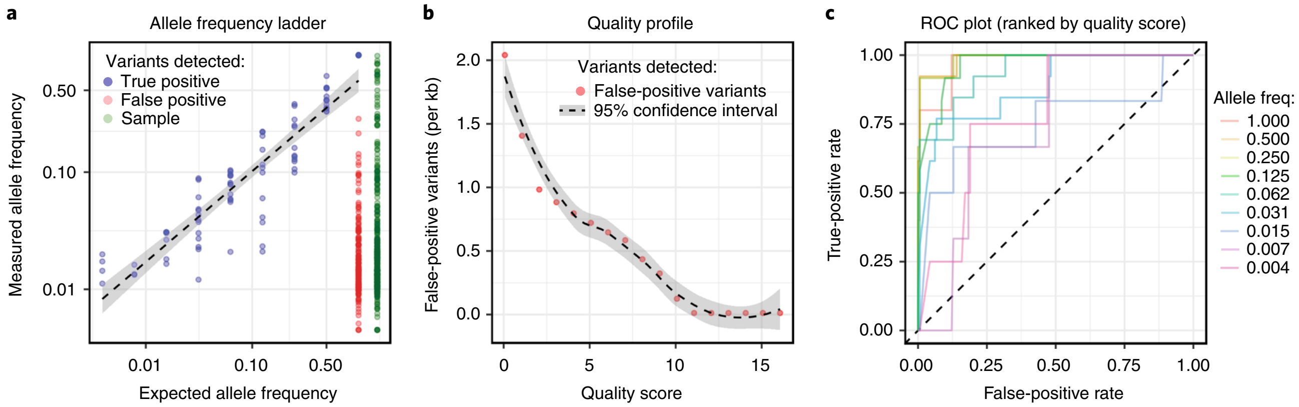


Fig. 11 | Performance evaluation of somatic variant calling by anaquin. a, Scatterplot indicating the observed and expected allele frequencies for variants encoded by sequins. Linear regression (black dashed line) and 95% confidence interval (shaded region) are indicated. The plot indicates the sensitivity and quantitative accuracy with which somatic variants are detected within the assay. **b**, Scatter plot illustrates the relationship between quality score and the detection of false-positive variants, and can be used to inform the imposition of quality-score thresholds that improve diagnostic specificity. **c**, Receiver operating characteristic (ROC) curve indicating true-positive rate (TPR) and false-positive rate (FPR) of called variants, as ranked by variant quality scores (somatic empirical variant scores (EVSs)). Expected allele frequency groups are plotted individually, with each curve indicating the diagnostic performance of somatic variant detection at a given frequency level.

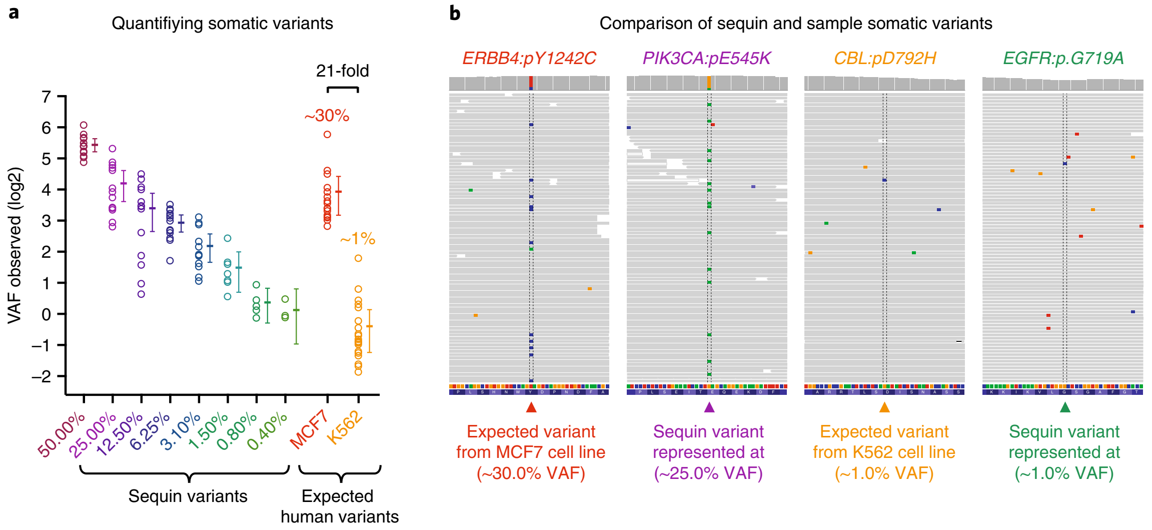


Fig. 12 | Comparison of expected human and sequin variants analyzed by targeted sequencing. a, Observed VAFs for variants annotated in the MCF7 and K562 cell lines, represented in the mock tumor mixture at ~30% and ~1% cellularity, respectively. Sequin VAF ladder provides an internal reference scale against which to assess the sensitivity and quantitative accuracy of candidate variants detected in the accompanying sample. Whiskers show median ± s.d. for each VAF level. **b**, Genome browser view (from IGV) shows sites of *ERBB4:pY1242C* and *CBL:pD792H* mutations in the MCF7 and K562 cell lines, respectively. Corresponding sequin variants represented at similar VAFs are also shown, providing matched examples to assist interpretation of candidate sample variants.

Given that sequin variants are represented at known VAFs, the quantitative accuracy and diagnostic performance at different VAF levels can be evaluated using anaquin ‘somatic’ (Step 25). Overall, the analysis of the tumor sequins indicates a strong correlation between observed and expected VAFs ($R^2 = 0.83$; Fig. 11). Although synthetic mutations could be detected at VAF levels as low as 0.4% (4/13 detected) within the assay, the measurement of low-VAF mutations nevertheless exhibits decreasing sensitivity and increasing quantitative variation.

Strelka2 also detected many erroneous variant candidates in sequin alignments, typically at low VAFs. To illustrate this, the anaquin ‘somatic’ tool generates an individual receiver operating characteristic (ROC) curve for each VAF level encoded within sequins, allowing us to assess the performance of variant detection for different frequencies (Fig. 11). This analysis shows that TP variants below ~3% VAF could not be reliably distinguished from FPs, resulting in a poor area under the curve (AUC) score of 0.84, and illustrates that the diagnostic limit is higher than the limit of detection for our targeted sequencing assay.

For comparison, we also evaluated the detection of variants (as annotated by the Cancer Cell Line Encyclopaedia[14]) harbored by the cell lines used in the mock tumor sample. Because MCF7 cells were included at higher cellularity (~30%) than K562 cells (~1%) in the tumor mixture, MCF7 variants were detected at a proportionally higher VAF than K562 variants (~21-fold observed difference), with the sequin VAF ladder providing a quantitative reference with which to assess this difference (Fig. 12a). As a result, MCF7 variants, such as *ERBB4:pY1242C*, can be confidently identified, whereas K562 variants, such as *CBL:pD792H*, cannot be reliably distinguished from FPs (Fig. 12b). This illustrates the concordance of annotated human variants with synthetic sequin variants at corresponding VAF levels and shows the utility of sequins for evaluating the diagnostic limits of our NGS assay.

A more detailed summary of these statistics, including information on variant types, contexts and allele frequencies, can be found within the somatic_report.html file from the somatic_out/ directory (Step 25). See Supplementary Note 2 to access an example report.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All next-generation sequencing libraries and associated data files, including synthetic sequences and variant annotations, are available for download at http://www.sequinstandards.com/resources/#nature_protocols. Please see the 'Equipment setup' section and Supplementary Notes 1 and 2 for further details.

Code availability

Anaquin source code is available from https://github.com/sequinstandards/RAnaquin.

References

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
2. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
3. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752–756 (2017).
4. Goldfeder, R. L. et al. Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 24 (2016).
5. Ross, M. G. et al. Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
6. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
7. Clark, M. J. et al. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
8. Lam, H. Y. K. et al. Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2011).
9. Gargis, A. S. et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat. Biotechnol.* **30**, 1033–1036 (2012).
10. Deveson, I. W. et al. Chiral DNA sequences as commutable controls for clinical genomics. *Nat. Commun.* **10**, 1342 (2019).
11. Deveson, I. W. et al. Representing genetic variation with synthetic DNA standards. *Nat. Methods* **13**, 784–791 (2016).
12. Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. *Nat. Rev. Genet.* **18**, 473–484 (2017).
13. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
14. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
16. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
17. Kim, S. et al. Strelka2: fast and accurate variant calling for clinical sequencing applications. *Nat. Methods* **15**, 591–594 (2018).

18. Wong, T., Deveson, I. W., Hardwick, S. A. & Mercer, T. R. ANAQUIN: a software toolkit for the analysis of spike-in controls for next generation sequencing. *Bioinformatics* **33**, 1723–1724 (2017).
19. Hodges, E. et al. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
20. Albert, T. J. et al. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
21. Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
22. Hardwick, S. A. et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat. Commun.* **9**, 3096 (2018).
23. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
24. Zook, J. M. & Salit, M. Genomes in a bottle: creating standard reference materials for genomic variation—why, what and how?. *Genome Biol.* **12**, P31 (2011).
25. Sims, D. J. et al. Plasmid-based materials as multiplex quality controls and calibrators for clinical next-generation sequencing assays. *J. Mol. Diagn.* **18**, 336–349 (2016).
26. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
27. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
28. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
29. Zheng, G. X. Y. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
30. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
31. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
32. Kavak, P. et al. Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics* **33**, i161–i169 (2017).
33. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
34. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
35. Murphy, K. M. et al. Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. *J. Mol. Diagn.* **8**, 305–311 (2006).
36. Ka, S. et al. HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics* **18**, 258 (2017).
37. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

Acknowledgements

The authors would like to thank the following funding sources: Australian National Health and Medical Research Council (NHMRC) Australia Fellowships (1062470 to T.R.M.), APP1108254 (to B.S.K.) and APP1114016 (to J.B). I.W.D is supported by a Cancer Institute NSW Early Career Fellowship (2018/ECF013). T.R.M. and T.W. are supported by a Paramor Family Fellowship. S.A.H. is supported by an Australian Postgraduate Award scholarship. A.L.M.R. is supported by a University of New South Wales Sydney Tuition Fee Scholarship. The contents of the published material are solely the responsibility of the administering institution, a participating institution or individual authors and do not reflect the views of the NHMRC.

Author contributions

J.B., B.S.K. and C.B. contributed materials. J.B. performed the experiments. T.W., I.W.D., S.A.H. and A.L.M.R. carried out the bioinformatic analysis. J.B., T.W., I.W.D. and T.R.M. wrote the manuscript. All authors conceived the study and contributed to manuscript preparation.

Competing interests

The authors declare competing interests: the Garvan Institute of Medical Research has filed patents covering aspects of sequencing controls.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-019-0175-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to I.W.D. or T.R.M.

Journal peer review information: *Nature Protocols* thanks Justin Zook and other anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 30 June 2017; Accepted: 3 April 2019;
Published online: 19 June 2019

Related links

Key references using this protocol

Deveson, I. W. et al. *Nat. Commun.* **10**, 1342 (2019): https://doi.org/10.1038/s41467-019-09272-0

Hardwick, S. A., Deveson, I. W. & Mercer, T. R. *Nat. Rev. Genet.* **18**, 473–484 (2017): https://doi.org/10.1038/nrg.2017.44

Deveson, I. W. et al. *Nat. Methods* **13**, 784–791 (2016): https://doi.org/10.1038/nmeth.3957 natureresearch

Corresponding author(s): Dr Tim Mercer
Dr Ira Deveson

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Not applicable
Data analysis	anaquin (s3.amazonaws.com/sequins/nature_protocols/anaquin_1.0.0.zip) BWA (v0.7.16) SAMtools (v1.9) Strelka2 (v2.9.2) GATK (v4.0.0) Trim Galore (v0.5.0) IGV (v2.4.7) R (v3.4.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability Statement
All next-generation sequencing libraries and associated data files, including synthetic sequences and variant annotations, are available for download at www.sequinstandards.com/resources/#nature_protocols. Please see the EQUIPMENT SETUP section and Supplementary Notes 1 and 2 for further details.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Single samples only are analzed in this protocol
Data exclusions	No data is excluded, except where described in the protocol
Replication	Single samples only are analzed in this protocol
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

All the cell lines utilised in the manuscript represent genomically well-characterised samples of cancer patient and normal individuals. Furthermore, none of these cell lines feature in the Database of Cross-Contamination of Misidentified Cell Lines, V8.
We have included the following comment within the manuscript (Section 2.1.1):
"To ensure purity of the sample materials, cell lines used in the WGS and targeted sequencing demonstrations were obtained from the American Type Culture Collection (MCF7, K562) and Coriell Institute for Medical Research (NA12878), cultured in a dedicated facility and DNA extracted using standard methods (DNeasy Blood and Tissue Kit #69504, Qiagen)."

Authentication	See above
Mycoplasma contamination	See above
Commonly misidentified lines (See ICLAC register)	None