Representing genetic variation with synthetic DNA standards

Ira W Deveson[1,2,5], Wendy Y Chen[1,3,5], Ted Wong[1], Simon A Hardwick[1,3], Stacey B Andersen[4], Lars K Nielsen[4], John S Mattick[1,3] & Tim R Mercer[1,3]

The identification of genetic variation with next-generation sequencing is confounded by the complexity of the human genome sequence and by biases that arise during library preparation, sequencing and analysis. We have developed a set of synthetic DNA standards, termed 'sequins', that emulate human genetic features and constitute qualitative and quantitative spike-in controls for genome sequencing. Sequencing reads derived from sequins align exclusively to an artificial *in silico* reference chromosome, rather than the human reference genome, which allows them them to be partitioned for parallel analysis. Here we use this approach to represent common and clinically relevant genetic variation, ranging from single nucleotide variants to large structural rearrangements and copy-number variation. We validate the design and performance of sequin standards by comparison to examples in the NA12878 reference genome, and we demonstrate their utility during the detection and quantification of variants. We provide sequins as a standardized, quantitative resource against which human genetic variation can be measured and diagnostic performance assessed.

Next-generation sequencing (NGS) can be used to determine an individual's genome sequence and identify disease-associated mutations[1–3]. Accordingly, NGS has become a principal tool in biomedical research and clinical diagnostics[4]. However, numerous variables influence the sensitivity and precision of variant detection, including biases arising during amplification and library preparation, insufficient or heterogeneous sequencing coverage, and sequencing errors[1,5,6]. The complexity of the human genome sequence makes subsequent alignment and analysis challenging, with short reads often aligning ambiguously within repetitive or low-complexity sequences. Variable allele frequencies and diversity in the architecture of genetic variants, which range from single base substitutions to large-scale genomic rearrangements, further confound their accurate identification[1,5,6].

We have developed a set of synthetic sequencing spike-in standards (sequins) that emulate features of a human genome, such as sites of genetic variation. Sequins are synthetic DNA molecules, up to ~10 kb in length, that represent real genetic features of interest. The gross architecture, nucleotide composition and alignability of these features are maintained, but sequins are entirely distinct in primary sequence. Therefore, sequin sequences do not align to any natural genome but instead to an artificial *in silico* reference chromosome, which contains all the features of a real human chromosome.

Sequins provide the physical representation of salient regions or features within this artificial reference chromosome. Sequins can be added to a DNA sample before library preparation and undergo concurrent sequencing and analysis, with derivative reads aligning exclusively to the *in silico* chromosome and, in this way, being partitioned from the accompanying genetic sample for parallel analysis as internal controls (**Fig. 1** and **Supplementary Fig. 1**).

Sequins can be used as qualitative controls to assess multiple stages of the NGS workflow, inform bioinformatic analyses and establish an internal reference scale against which quantitative features of genome biology can be measured (**Fig. 1** and **Supplementary Fig. 1**). Here we describe the use of sequins to emulate common and disease-associated genetic variation, including single nucleotide variants (SNVs) and small insertions and deletions (indels), as well as large structural variants and copy-number variation; and we demonstrate the utility of sequins in genome sequencing experiments.

RESULTS

Design and preparation of DNA sequins

We first designed an ~11-Mb artificial *in silico* chromosome (*chrIS_D*) by proportionately sampling regions from the human genome (Hg38)[7] according to size, GC content and repeat density (**Fig. 1** and **Supplementary Fig. 2**). Sequences were inverted (i.e., 3′ to 5′) and, in local regions where homology to natural genome sequences remained, additional shuffling and/or nucleotide substitution was performed (**Supplementary Fig. 2**; see Online Methods).

A proportional representation of common genetic variation within human populations was then encoded within the *chrIS_D* sequence (**Fig. 1**). First, we ranked common human genetic variation (dbSNP 141; ref. 8) according to variation type, size and

[1]Genomics and Epigenetics Division, Garvan Institute of Medical Research, New South Wales, Australia. [2]School of Biotechnology and Biomolecular Sciences, Faculty of Science, The University of New South Wales, New South Wales, Australia. [3]St. Vincents Clinical School, Faculty of Medicine, The University of New South Wales, New South Wales, Australia. [4]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Queensland, Australia. [5]These authors contributed equally to this work. Correspondence should be addressed to T.R.M. (t.mercer@garvan.org.au).

RECEIVED 23 MARCH; ACCEPTED 28 JUNE; PUBLISHED ONLINE 8 AUGUST 2016; DOI:10.1038/NMETH.3957

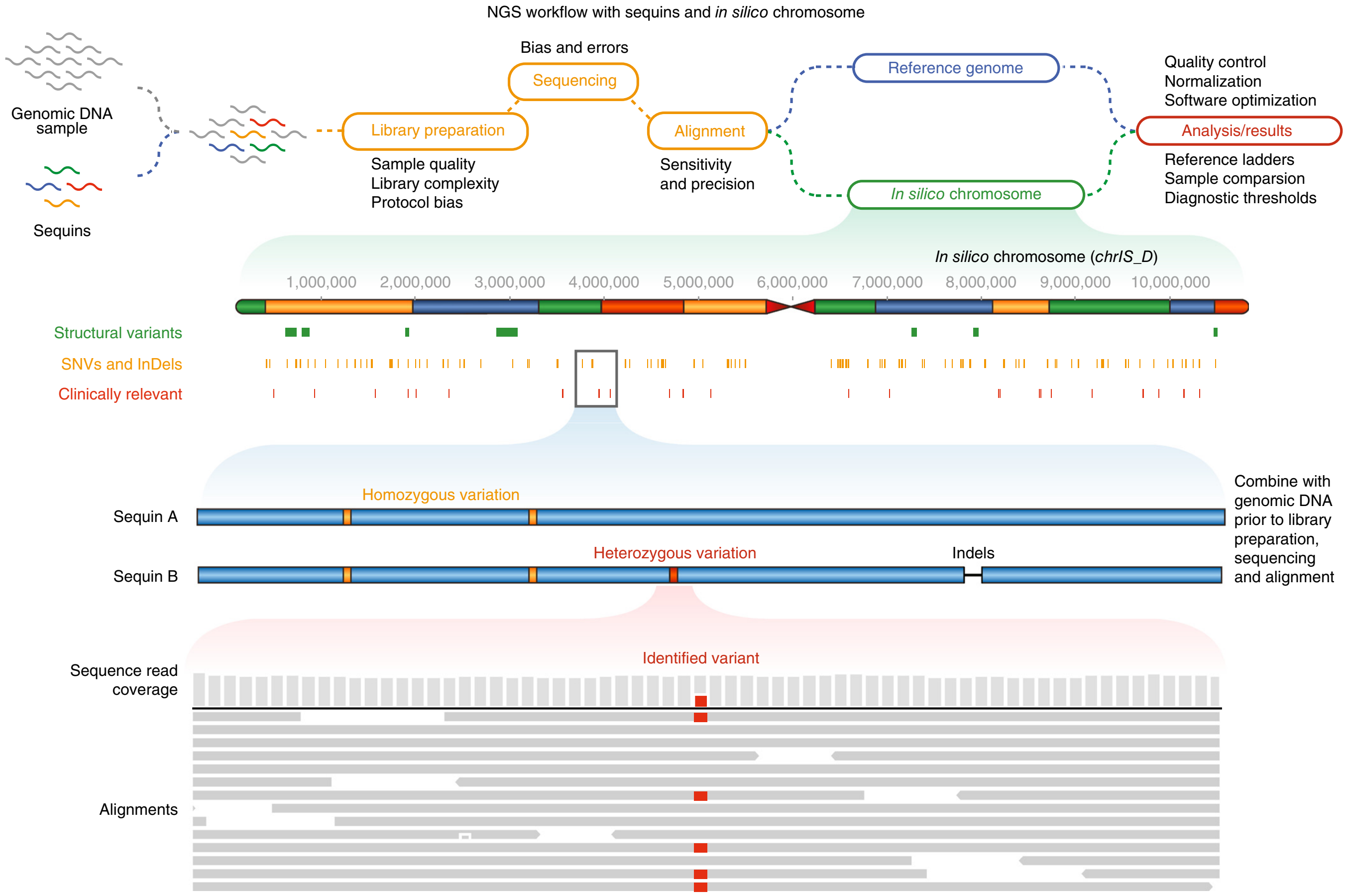


Figure 1 | Sequin standards representing genetic variation. Sequins are added to a DNA sample before library preparation, sequencing and analysis (yellow). Reads from the DNA sample align to the reference genome (blue; e.g., Hg38), while sequin reads align to an artificial *in silico* chromosome (green). Sequins can be used to assess many stages of the NGS workflow; examples are indicated (red). The artificial *in silico* chromosome (*chrIS_D*) contains encoded genetic variation, including common (yellow) and disease-associated (red) SNVs and indels, as well as large structural variants (green). Sites of synthetic genetic variation are represented by paired DNA sequins. Each sequin represents one allele of a diploid genotype, with *chrIS_D* constituting a consensus reference sequence, analogous to Hg38. Homozygous (yellow) and heterozygous (red) genotypes can be emulated by manipulating the abundance ratio of reference to variant standards within a pair. The lower panel shows sequin alignments (gray) identifying a synthetic heterozygous variant site with respect to *chrIS_D*.

nucleotide content; and we systematically sampled 223 SNVs and 176 indels (**Supplementary Fig. 3b–d**; see Online Methods). Selected variants and their surrounding native genome sequences were incorporated into *chrIS_D*. We were unable to emulate a small subset of variants that reside in highly repetitive and/or low-complexity sequences because of both synthesis difficulties and the high risk of cross-alignment to the human genome that these confer (**Supplementary Fig. 3e**).

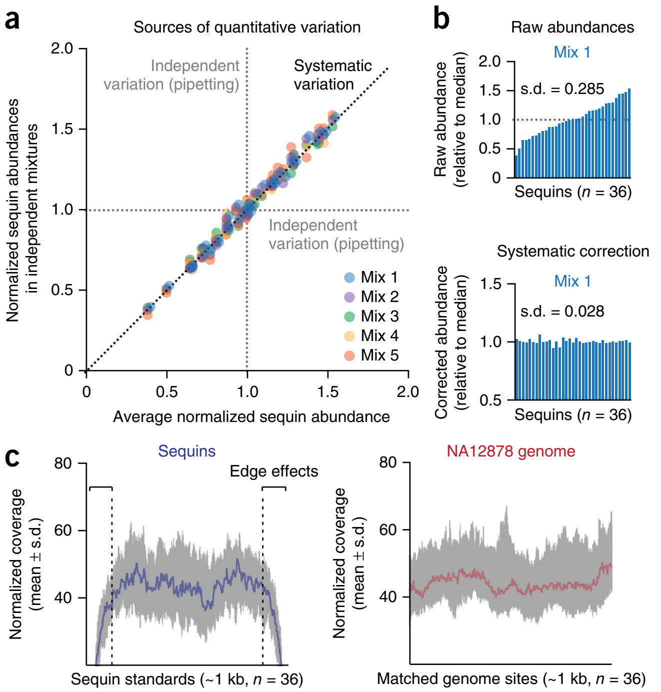
While the artificial chromosome remained *in silico*, sequins were next designed to represent sites of genetic variation with respect to this reference sequence. A pair of sequins was used to represent the two alleles of a diploid genotype, with *chrIS_D* constituting a consensus reference sequence analogous to Hg38 (ref. 7) (**Fig. 1**). To emulate a heterozygous variant, for example, one sequin was designed to represent the reference allele and another to represent the variant allele (**Fig. 1**). In total, 36 pairs of sequin standards were designed, collectively encoding 167 homozygous and 245 heterozygous instances of variation with respect to *chrIS_D*. Sequins were ~1 kb in length, including ~150-bp flanking regions included to prevent low coverage of terminal variants due to sequencing 'edge effects'[9].

To validate their design, we initially simulated paired-end sequencing libraries[10] from sequins and from the human genome. Simulated reads were aligned to a combined *chrIS_D* and Hg38 index using BWA-mem[11]. Sequin-derived reads aligned exclusively to *chrIS_D*, and human reads aligned exclusively to Hg38. To force cross-alignment, sequins reads were required to be aligned to Hg38 in the absence of *chrIS_D* and with a reduced mismatch penalty (which also resulted in widespread misalignment of simulated human reads; **Supplementary Fig. 4a**).

DNA sequins were synthesized and hosted within template vectors, from which they could be prepared by restriction digestion. We combined all sequin pairs at equimolar concentration and sequenced this 'flat mixture' using standard protocols (see Online Methods). As with simulated reads, we observed no sequin-derived reads aligning to the Hg38 and, conversely, no reads from whole-genome sequencing of individual NA12878 (ref. 12) aligning to *chrIS_D* (**Supplementary Fig. 4b**). This confirmed that natural and synthetic sequences can be safely partitioned according to their alignment.

We next investigated the magnitude and direction of variability in the measured abundances of individual sequin standards.

Figure 2 | Assessing quantitative variability within and between sequin mixtures. Mix, mixture. (**a**) Comparison of observed abundances (median per-base coverage) of individual sequin standards (normalized to median abundance within a mixture) between five independent mixtures with sequins added at equimolar concentrations. The vector directions associated with systematic and independent variation are indicated. (**b**) Histograms indicate observed abundances (normalized to median) of sequin standards in one mixture before and after (matched order) correction for systematic biases. (**c**) Plots show mean coverage (± s.d.) across the averaged length of sequin standards (left, *n* = 36) in comparison to coverage across matched ~1-kb windows in the NA12878 genome (right, *n* = 36) after coverage of sequins is calibrated to accompanying human genome. Terminal regions (~150 bp) of sequins where edge effects impact coverage are indicated (dashed lines) and were excluded during calibration and downstream analysis.



These may arise from (i) intrinsic sequence-specific biases during library preparation, sequencing and alignment; and (ii) aliquot variability during the preparation of sequin mixtures. To distinguish these sources of variation, we prepared and sequenced five independent sequin flat mixtures. We assumed that sequence-specific variation impacts individual sequins in a reproducible manner in different mixtures, whereas aliquot variation is independent of standard identity and random between mixtures (**Fig. 2a** and **Supplementary Fig. 5**). We found that the majority of variability in the measured abundances of individual sequins within a mixture (average s.d. of abundances, 0.284) was reproducible, with the remaining variability (pipetting variation) constituting <10% of total variation (s.d. after systematic bias correction, 0.023, **Fig. 2b**; **Supplementary Fig. 5**). The measurement of independent variation can inform further improvements in mixture preparation, while the measurement of reproducible variation provides an assessment of the sequence-specific biases that impact NGS.

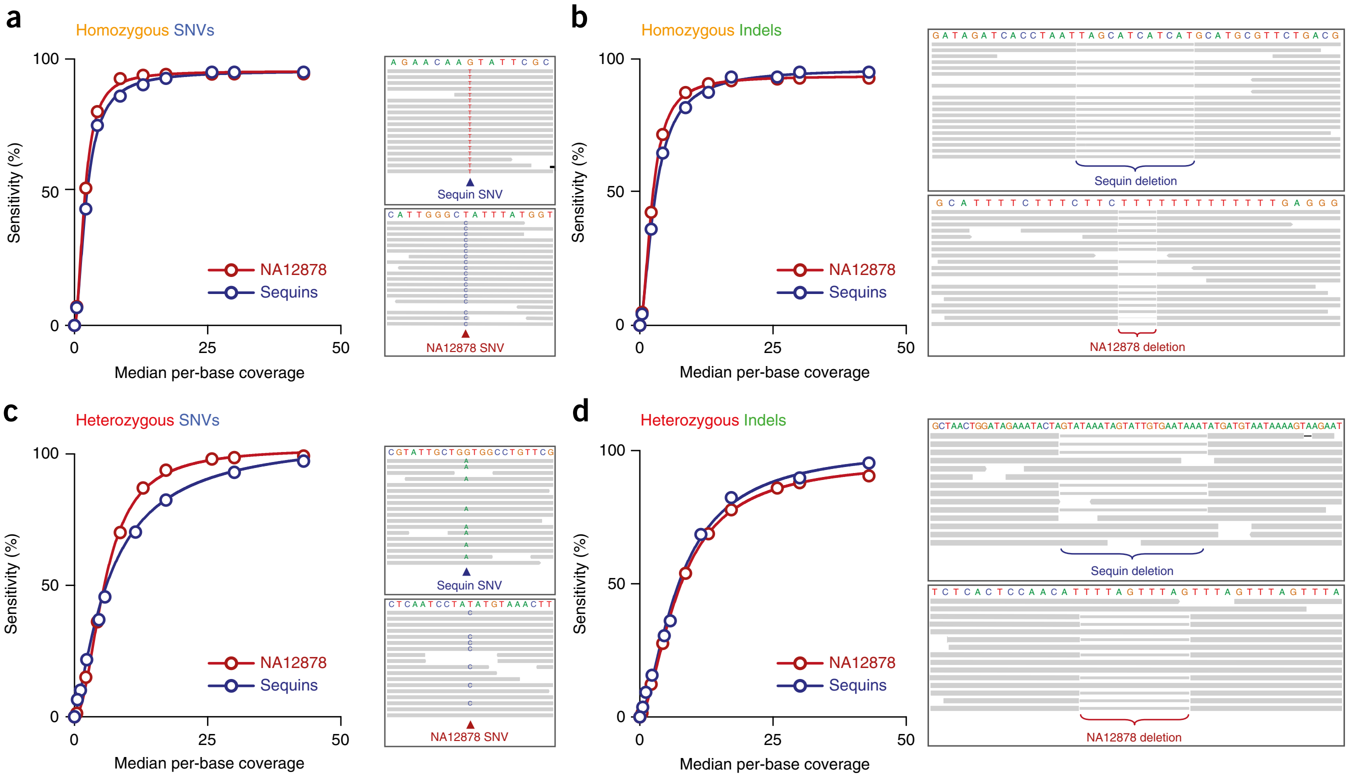


Figure 3 | Validation of DNA sequins against the NA12878 reference genome. (**a**–**d**) Cumulative population distributions describe the sensitivity of detection for synthetic variation represented with sequins (blue) and natural genetic variation annotated in NA12878 genome (red) as a function of increasing sequencing depth (median per-base coverage). Plots depict homozygous SNVs (**a**) and indels (**b**) and heterozygous SNVs (**c**) and indels (**d**) alongside accompanying genome browser examples (right panels), showing sequence alignments supporting variant identification.

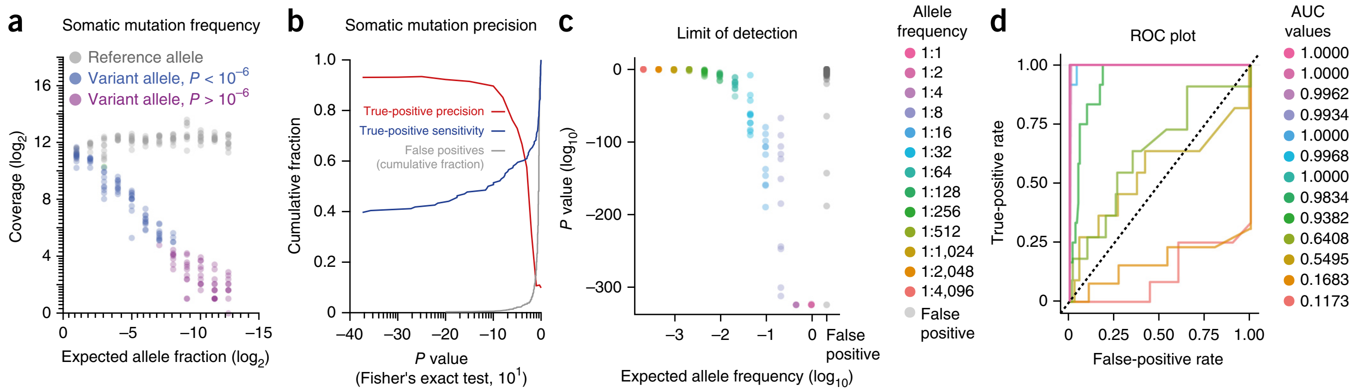


Figure 4 | Using sequins to measure somatic variant-allele frequency. (**a**) Read counts supporting reference (gray) and variant alleles (blue and purple) for synthetic SNVs are plotted against expected variant-allele fractions. SNVs in blue were called with a minimum confidence of $P < 10^{-6}$ (Fisher's exact test), while green variants failed to surpass this threshold. (**b**) Frequency distributions describe the sensitivity (blue) and precision (red) of synthetic SNV detection and the accumulation of false-positive calls (gray) as a function of variant calling confidence threshold (Fisher's exact test). (**c**) Limit-of-detection plot indicates the confidence with which synthetic SNVs were detected relative to their expected allele frequencies. False-positive SNV calls (gray) are included for comparison. (**d**) Receiver operating characteristic (ROC) curve illustrates the true-positive relative to false-positive detection rate for synthetic SNVs when ranked according to variant confidence (Fisher's exact test on reference to variant reads). Each allelic fraction is plotted independently with corresponding AUC values providing a measure of diagnostic performance.

Genetic variation

To demonstrate the utility of sequins in genome sequencing, we added a sequin flat mixture to NA12878 genomic DNA (1% final concentration). This combined sample was subject to library preparation and sequencing, and the resulting reads were aligned to a combined Hg38 and *chrIS_D* index. Sequin-derived reads (aligning to *chrIS_D*) were indistinguishable from reads from the NA12878 genome (aligning to Hg38) in quality, and they exhibited equivalent mapping confidence between the artificial and human chromosomes (**Supplementary Fig. 6a–c**).

Alignments were subsampled to achieve overall sequencing coverage equivalent to the accompanying NA12878 genome (**Fig. 2c** and **Supplementary Fig. 7a**; see Online Methods). Following subsampling, sequins exhibited equivalent depth (43-fold per base median) and heterogeneity (19-fold interquartile range) to matched sites in the human genome (**Fig. 2c** and **Supplementary Fig. 7a**,**b**). Such an approach allows the sequencing coverage of sequins to be precisely calibrated to any genome regions of interest, regardless of the fractional concentration at which they were added.

We next compared the identification of synthetic sequin variation to the identification of high-confidence bona fide variants in the NA12878 human genome (Platinum Genome Project). Variants were identified with respect to a combined *chrIS_D* and Hg38 index using the Genome Analysis Toolkit (GATK) best-practice pipeline[13–15]. At maximum matched depth (43-fold median-per-base coverage), we identified 95% and 99% of heterozygous synthetic and NA12878 SNVs, respectively, and 99% of both synthetic and NA12878 homozygous SNVs (**Fig. 3a**,**c**). Similarly, we found 95% and 93% of heterozygous synthetic and NA12878 indels, respectively, and 100% and 98% of homozygous synthetic and NA12878 indels (**Fig. 3b**,**d**). We identified three false-positive indels and no false-positive SNVs on *chrIS_D* at recommended hard-filtering thresholds[13–15]. By progressively subsampling alignments to both *chrIS_D* and the Hg38 genome, we confirmed that depreciating sequencing coverage has a comparable impact on the detection of synthetic and NA12878 variants (**Fig. 3a–d**). This demonstrates that variants represented by sequins perform analogously to bona fide variants within the well-characterized NA12878 genome, verifying their suitability as an internal positive control set for genome sequencing.

Somatic mutations

Somatic mutations implicated in cancer can occur at a range of allele frequencies on account of copy-number variation, tumor heterogeneity and sample impurity[16–21]. The combination of sequin pairs at different relative concentrations can be used to establish a reference scale against which to measure allele frequency. We titrated variant allele standards at a two-fold serial dilution relative to their corresponding reference allele standards, generating a scale of allele frequency from 1:1 (heterozygous) to 1:4,096 (**Supplementary Fig. 8a**), with each fraction represented by three pairs. This staggered mixture was added to NA12878 genomic DNA, and the combined sample was sequenced, with variants subsequently identified and quantified using VarScan2 (ref. 22). We used this sequin mixture as an internal standard against which to measure the sensitivity (*sn*) and precision (*pr*) of somatic mutation detection.

We added sequins at high initial concentration (10%) to achieve saturating sequencing coverage for a quantitative analysis of sequins. At full depth (>25,000-fold coverage), we obtained at least one supporting read for all except two synthetic SNVs, both of which belonged to the lowest allelic fraction (1:4,096; **Fig. 4a**). We also observed a strong relationship between the variant read-count fraction and the expected allele frequency for SNVs and indels (R^2 = 0.88 and 0.45; **Supplementary Fig. 8b**,**c**). However, in real experimental scenarios, sequencing depth imposes a limit on the sensitivity of mutation detection. To illustrate this, we subsampled alignments to simulate different sequencing depths and measured the sensitivity of variant detection across this range (**Supplementary Fig. 9a–d**). For example, 100-fold sequencing coverage was insufficient to detect the majority of variants present at an allele frequency or below 12.5% (**Supplementary Fig. 9a**,**b**).

The erroneous detection of false-positive variants in sequins relative to *chrIS_D* allowed us to assess the precision of variant

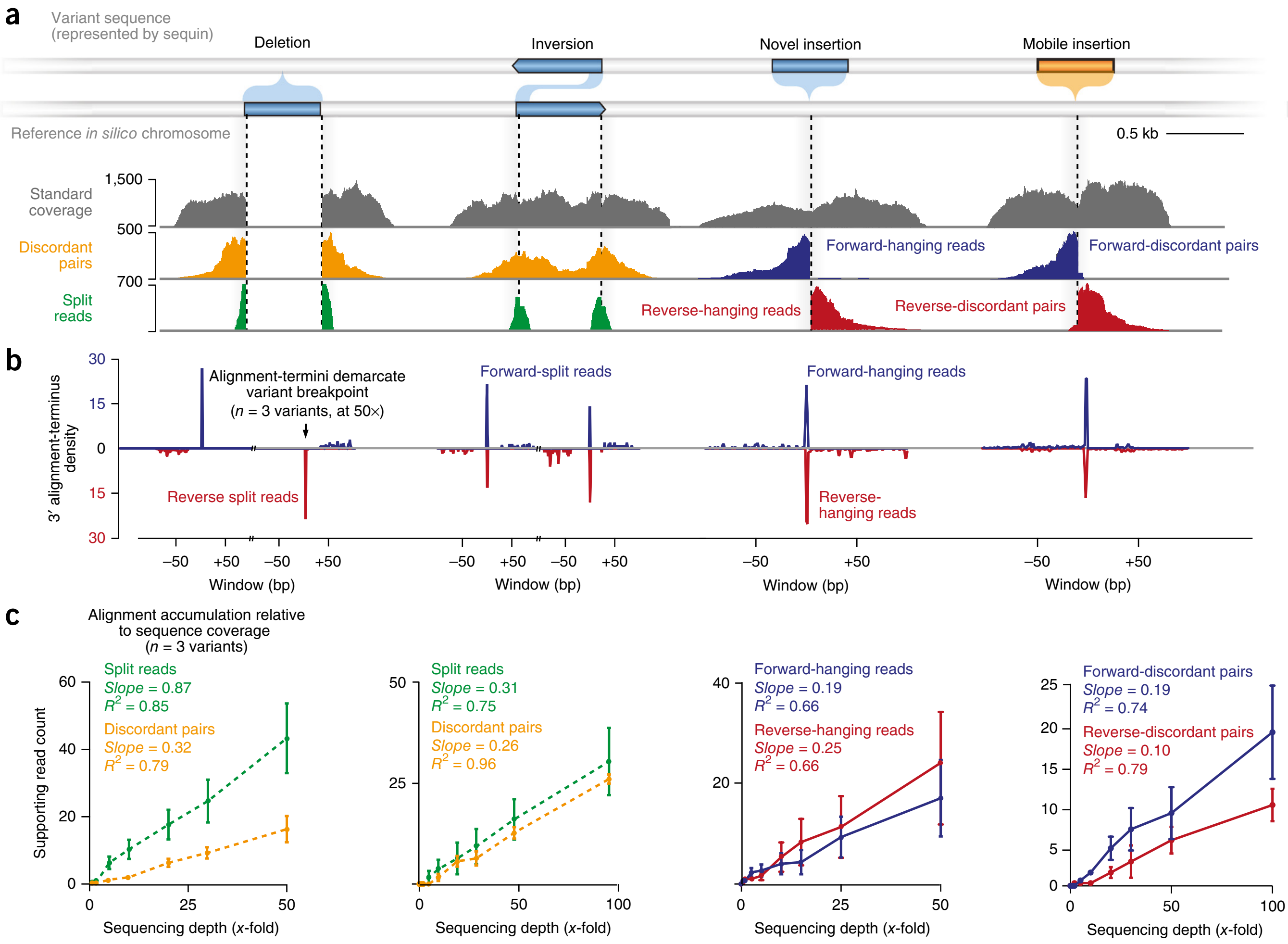


Figure 5 | Representing structural variation with DNA sequins. (**a**) DNA sequins were designed to represent synthetic deletions ($n = 3$), inversions ($n = 3$), novel sequence insertions ($n = 3$) and mobile-element insertions ($n = 3$) with respect to *chrIS_D*. Examples illustrate characteristic alignment profiles for each structural variant class. (**b**) Normalized density of 3′ alignment termini, in forward (blue) and reverse (red) orientation, is plotted with respect to synthetic variant breakpoint sites (mean, $n = 3$ for each class). (**c**) The accumulation of split-read (green) or discordant-pair (yellow) reads is plotted against sequencing depth (*x*-fold coverage). Mean indicated, error bars = s.d., $n = 3$. Discordant pairs and hanging reads are shown in forward and reverse orientation for insertions.

detection. At full coverage, where sensitivity is maximized, 1,254 potential erroneous SNVs and 64 erroneous indels were identified in *chrIS_D* (**Fig. 4b** and **Supplementary Fig. 10f**). These erroneous calls resulted primarily from sequencing errors, which occurred at low frequency, and these calls must be filtered to optimize precision. Sequins can be used to evaluate different filtering strategies—for instance, supporting read counts, allele frequency, the ratio of supporting reads on each strand, the quality of variant-supporting nucleotide calls and variant confidence scores[22]—for their ability to eliminate false-positive variant calls (**Supplementary Fig. 10a–d**).

To illustrate one such approach in detail, we ranked true- and false-positive variants sites according to variant confidence scores[22] (*P* values; Fisher's exact test on supporting read counts) to generate receiver operating characteristic (ROC) curves that define the precision of variant detection at each expected allelic fraction (**Fig. 4c,d** and **Supplementary Fig. 10**). For example, at allele frequencies below 1/128 (at *sn* = 75%, *pr* = 94%; **Fig. 4c,d**), true-positive SNVs cannot be confidently distinguished from false-positive calls (at 1/256, *sn* = 75%, *pr* = 54%; **Fig. 4c,d**). Internally generated and empirically derived assessments enable a user to balance sensitivity and precision according to experimental or clinical requirements.

Structural variation

Structural variants (SVs), which encompass the deletion, insertion, translocation, inversion or rearrangement of large (>200-bp) regions of the human genome, are a major source of genetic variation and are associated with a range of diseases[23,24]. However, SVs are often difficult to detect with short-read sequencing due to their large size, complex architecture and incorporation of low-complexity or repetitive sequences[23,25]. Methods for detecting SVs use evidence from discordant read-pair alignment, split-read alignment and read-depth anomalies[23,25,26].

We designed sequin standards to represent deletions ($n = 3$), inversions ($n = 3$), novel sequence insertions ($n = 3$) and mobile element insertions ($n = 3$) with respect to *chrIS_D* (**Fig. 5a**). Representative SV sequences selected from the Database of Genome Variants[27] were inverted and/or shuffled to abolish homology while maintaining their original architecture, nucleotide

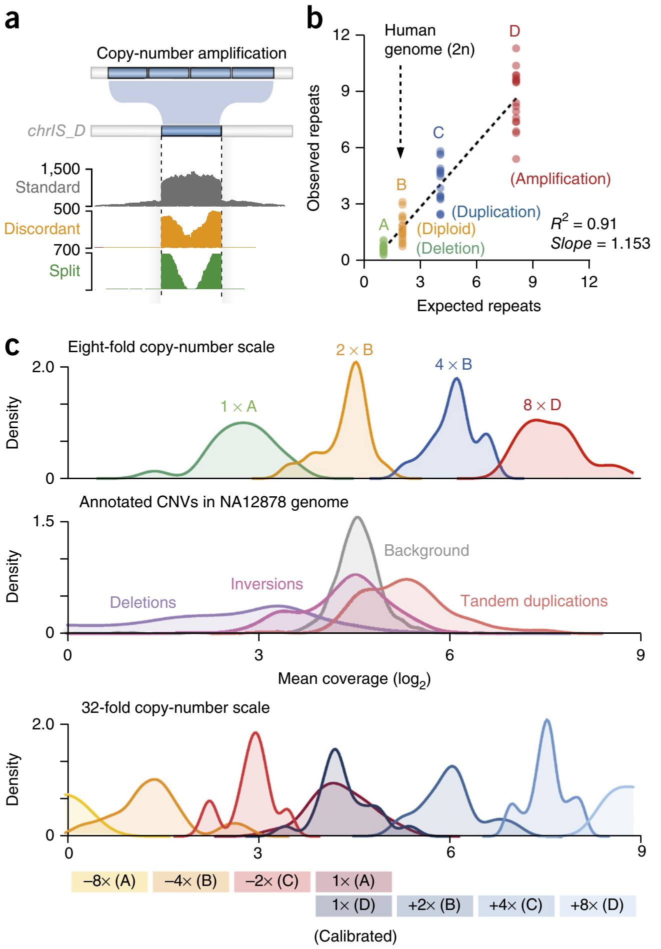


Figure 6 | Representing copy-number variation and repeat DNA with sequins. (**a**) DNA sequins designed to represent sites of copy-number amplification (n = 3) with respect to *chrIS_D*. (**b**) Scatter plot indicates observed relative to expected repeat copies for variant DNA elements. (**c**) Frequency distribution illustrates the mean alignment coverage across copy-number elements that constituted the reference scale (upper panel). Frequency distribution illustrates the alignment coverage across annotated CNVs in the NA12878 genome for comparison to copy-number sequin scale (middle panel). Frequency distribution illustrates that the calibration of alternate copy-number elements enables extension of the scale above and below the genome mean, and extends to 32-fold scale (lower panel).

composition and repetitiveness (see Online Methods). These were incorporated into *chrIS_D*, and sequin standards were designed to represent each synthetic SV site (**Fig. 5a**). Sequins were manufactured, combined into a flat mixture, sequenced and aligned to *chrIS_D*. Following alignment, we used the Lumpy package[28] to detect synthetic SVs based on read depth and discordant-pair and split-read alignments (**Fig. 5** and **Supplementary Figs. 11–14**).

Synthetic deletions and inversions could be confidently resolved at 20-fold and 30-fold sequencing coverage, respectively (**Fig. 5b–d**). At inversion sites, split-read (slope = 0.31, R^2 = 0.75) and discordant-pair (slope = 0.26, R^2 = 0.96) alignments accumulated linearly and at comparable rates with increasing coverage (**Fig. 5d**), whilst at deletion sites, split-read alignments accumulated ~2-fold faster (slope = 0.87, R^2 = 0.85; **Fig. 5d**) than discordant pairs (slope = 0.32, R^2 = 0.79; **Fig. 5d**). This alignment signature was also observed at high-confidence deletions annotated in the NA12878 human genome (**Supplementary Fig. 11**). Furthermore, the left and right breakpoint nucleotides of deletions and inversions were clearly demarcated by split-read alignment termini, with opposing alignment polarity distinguishing these two SV types (**Fig. 5c**).

Synthetic novel sequence insertions (~600 bp) exceeded read length and therefore did not generate split-read alignments. Instead, we observed an accumulation of discordant pairs, in which one mate cannot be aligned to the reference, at insertion sites (**Fig. 5b**). These 'hanging' alignments are polarized, with their 3′ termini converging on the insertion breakpoint from forward and reverse directions, accumulating linearly with increasing sequencing depth (slope = 0.19, 0.25, R^2 = 0.66, 0.25; **Fig. 5c,d**). This strand-specific profile provides a distinct and characteristic signal for novel sequence insertions that is similar to that observed at high-confidence insertions annotated in the NA12878 genome (**Supplementary Fig. 12**).

Transposable elements are a major component of the human genome, and active transposons are a source of human genetic variation[29]. Synthetic mobile elements were included in the artificial *in silico* chromosome design (see Online Methods), and sequins were constructed to represent examples of mobile element insertion events at nonannotated sites (n = 3; **Supplementary Fig. 13**). Mobile element insertions exhibited characteristic discordant-pair alignment signatures, with multi-mapped (mapping quality mapQ = 0) reads anchored to confident alignment partners (mapQ > 10) converging on each insertion site (**Fig. 5b–d** and **Supplementary Fig. 13**).

Copy-number variation

The duplication or deletion of large genome regions results in copy-number variation (CNV) between individuals and in disease[30]. We designed DNA sequins that represent CNVs of different size and copy number (2 copies × 621 bp, 4 copies × 202 bp and 6 copies × 96 bp) with respect to *chrIS_D* (**Fig. 6a** and **Supplementary Fig. 14**). Lumpy[28] successfully identified the two larger CNVs, using split-read and discordant-pair alignments (**Fig. 6a**), but failed to identify the smallest duplicated region (96 bp). At each site, we observed a mean read-depth shift between amplified and flanking regions that was proportionate to copy number (slope ratio = 5.47, 4.62 and 1.75, respectively, for 6-copy, 4-copy and 2-copy CNVs; **Supplementary Fig. 14b**).

Read-depth shifts can be used to identify and quantify CNVs in the human genome, although the accuracy of this approach is dependent on the depth and homogeneity of sequencing coverage[31]. We designed a set of sequins that form a quantitative scale against which to measure DNA copy number. Each individual sequin (n = 18) comprised four distinct 600-bp DNA elements incorporated at 1, 2, 4 and 8 copies, respectively (**Fig. 6b**). As this quantitative scale was established based on the copy number of conjoined DNA elements, rather than their input concentrations, this ensured that fold differences between elements were unaffected by aliquot variability during mixture preparation.

Following the construction and sequencing of CNV sequins, read depth was calibrated so that DNA elements present at two copies were matched to the accompanying diploid NA12878 human genome (**Fig. 6b**; see Online Methods). We observed a strong linear relationship between expected abundance and read depth across this quantitative scale (R^2 = 0.81, slope = 1.15; **Fig. 6b**), which can be used to assess the resolution with which copy-number integers are calculated via the mean-shift approach (**Fig. 6c** and **Supplementary Fig. 15a**). Given that any copy-number element can be calibrated to the genome, the scale can be extended to encompass up to 32-fold duplications relative to the genome (**Fig. 6c**).

We used this scale to assess the detection of CNVs previously annotated in the accompanying NA12878 genome[28,32], finding that 78.7% of deletions concord with one-copy deletion standards (by comparison to 2.7% of background controls; **Fig. 6c** and **Supplementary Fig. 15b**). The resolution of duplications was more challenging, with ~52.8% of annotated tandem duplications concurring with copy scale (by comparison to 3.5% of background controls; **Fig. 6c**), and lower confidence ascribed to the mean shift increase (**Supplementary Fig. 5b**).

A range of repeat DNA features, such as satellite DNA, rDNA, and viral or mitochondrial sequences, occur at tens to hundreds of copies in the human genome. To encompass this diversity, copy-number sequins were next combined across a two-fold dilution series to establish a 10^6-fold reference scale, in which the abundance of each DNA element is the product of its copy number within a sequin and the input concentration of that sequin (**Supplementary Fig. 15c**).

This extended DNA ladder was added to NA12878 genomic DNA (8% fractional concentration) and sequenced. Measured abundance (median coverage) of DNA elements scaled linearly with input concentration (amol μL^{-1}), down to a distinct inflection point at 4.88 amol μL^{-1} (piecewise linear regression; **Supplementary Fig. 15c**). This point represents the limit of quantification (LoQ)[33], below which the measurement of DNA abundance becomes increasingly variable and nonlinear (**Supplementary Fig. 15c**). The strong linear relationship above the LoQ ($R^2 = 0.97$; **Supplementary Fig. 15c**) is indicative of high quantitative accuracy, with sequins in this range constituting an internal and generic ladder against which a diversity of DNA repeat features could foreseeably be measured. To illustrate the landscape of DNA repeat features within the human genome, we have plotted several examples (**Supplementary Fig. 15c**), such as the mitochondrial chromosome, the rRNA operon, the Epstein–Barr virus genome and a range of satellite repeats, all of whose measured copy numbers were consistent with previous reports[34–37] against the sequin reference ladder.

DISCUSSION

Though NGS has emerged as a central tool in biomedical research, the development of controls with which to measure and mitigate the complex mixture of technical, informatic and biological variables that influence the outcomes of a given assay has been insufficient. Reference materials, including engineered cell lines and well-characterized individual genomes, act as valuable process controls[12]. However, while they recapitulate the size and complexity of the genome, such materials are not intended for direct addition to a sample and cannot be used to assess sample-specific variability or facilitate intersample normalization. With NGS already being applied in clinical contexts[38], there is a pressing need for internal standards that can ensure robust interpretation and analysis, enable comparison between multiple samples (including those handled by different laboratories) and provide a quantitative reference against which diagnostic thresholds may be anchored.

Sequin standards enjoy several advantages. First, sequins are added directly to the sample and, hence, constitute internal controls that enable both quantification of artifacts and variability during sample processing and intersample normalization. Second, the sequins can be mixed at precise ratios in order to generate reference scales with which to assess quantitative features of genome biology, such as allele frequencies. Finally, sequins can be used to simultaneously emulate bona fide genetic features and distinguish false-positive results, such as the erroneous detection of genetic variants within the artificial *in silico* chromosome. As a result, sequins enable empirical determination of both sensitivity and precision within a specific experiment, and they may be used to optimize performance for any assay that employs NGS.

The examples described here are intended as illustrative, rather than comprehensive, applications, and it should be borne in mind that almost any stage of an NGS workflow could foreseeably be interrogated with sequins. In an accompanying manuscript, Hardwick *et al.*[39] present a set of RNA sequins for use in transcriptome profiling by RNA-seq. Such versatility is a major advantage of sequin technology over existing spike-in approaches in DNA[40] or RNA sequencing[41], and synthetic sequin controls could, in principle, be designed to represent any feature of the genome or transcriptome. Sequins therefore constitute a distinct class of analytical tool, which promises to improve the breadth of features that can be confidently assessed with NGS.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing libraries have been deposited to the NCBI Short Read Archive (SRA) BioProject page under accession code PRJNA321434.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank our colleagues M. Cowley and M. Pinese for useful discussions and K. Ying for assistance in bioinformatic pipelines. The authors would like to thank the following funding sources: Australian National Health and Medical Research Council (NHMRC) Australia Fellowship (1062470 to T.R.M. and 1062606 to W.Y.C.). I.W.D. and S.A.H. are supported by Australian Postgraduate Award scholarships. The contents of the published material are solely the responsibility of the administering institution, a participating institution or individual authors, and they do not reflect the views of NHMRC.

AUTHOR CONTRIBUTIONS

T.R.M. conceived the project, designed sequins and synthetic chromosome and conceived experiments. W.Y.C. and S.B.A. prepared sequins and performed experiments. I.W.D., T.W. and T.R.M. performed data analysis. I.W.D., S.A.H., L.K.N., J.S.M. and T.R.M. prepared the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Goldstein, D.B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14**, 460–470 (2013).

2. Mwenifumbo, J.C. & Marra, M.A. Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14**, 321–332 (2013).
3. Gundry, M. & Vijg, J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutat. Res.* **729**, 1–15 (2012).
4. Katsanis, S.H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14**, 415–426 (2013).
5. Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
6. 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).
7. Rosenbloom, K.R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
8. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
9. Satya, R.V. & DiCarlo, J. Edge effects in calling variants from targeted amplicon sequencing. *BMC Genomics* **15**, 1073–1080 (2014).
10. Huang, W., Li, L., Myers, J.R. & Marth, G.T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
11. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
12. 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).
13. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 1–33 (2013).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
16. De Sousa E Melo, F., Vermeulen, L., Fessler, E. & Medema, J.P. Cancer heterogeneity—a multifaceted view. *EMBO Rep.* **14**, 686–695 (2013).
17. Meacham, C.E. & Morrison, S.J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
18. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
19. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).
20. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
21. Aran, D., Sirota, M. & Butte, A.J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
22. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
23. Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
24. Weckselblatt, B. & Rudd, M.K. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet.* **31**, 587–599 (2015).
25. Abel, H.J., Duncavage, E.J. & Duncavage, E.J. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet.* **206**, 432–440 (2013).
26. Pirooznia, M., Goes, F.S. & Zandi, P.P. Whole-genome CNV analysis: advances in computational approaches. *Front. Genet.* **6**, 138 (2015).
27. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L. & Scherer, S.W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
28. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
29. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
30. Zarrei, M., MacDonald, J.R., Merico, D. & Scherer, S.W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
31. Wineinger, N.E. *et al.* Statistical issues in the analysis of DNA copy number variations. *Int. J. Comput. Biol. Drug Des.* **1**, 368–395 (2008).
32. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. Preprint at *bioRxiv* http://dx.doi.org/10.1101/035493 (2015).
33. Armbruster, D.A. & Pry, T. Limit of blank, limit of detection and limit of quantitation. *Clin. Biochem. Rev.* **29** (Suppl. 1), S49–S52 (2008).
34. Gibbons, J.G., Branco, A.T., Yu, S. & Lemos, B. Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun.* **5**, 4850 (2014).
35. Lei, H. *et al.* Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. *BMC Genomics* **14**, 804 (2013).
36. Schaap, M. *et al.* Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics* **14**, 143 (2013).
37. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
38. Frampton, G.M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
39. Hardwick, S.A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* http://dx.doi.org/10.1038/nmeth.3958 (2016).
40. Lih, C.J. *et al.* Analytical validation and application of a targeted next-generation sequencing mutation-detection assay for use in treatment assignment in the NCI-MPACT trial. *J. Mol. Diagn.* **18**, 51–67 (2016).
41. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).

ONLINE METHODS

Design of artificial *in silico* chromosome. An *in silico* chromosome sequence was designed to represent features of real human chromosomes with artificial, nonhomologous primary sequences. We performed the following to assemble the *in silico* chromosome sequence (.fasta file).

Background sequence. To generate a background sequence into which synthetic variants could be incorporated, the human genome (Hg38) sequence was binned into 100-kb windows, which were ranked according to gene density, repeat density and GC content, then systematically sampled to ensure proportional representation.

Incorporating small-scale genetic variation (SNVs and indels). Our aim in selecting human variation was two-fold. First, we aimed to provide a proportional representation of human genetic variation in populations. To achieve this we initially ranked dbSNPs variants by type (SNV, insertion, deletion, other) and systematically selected each *n*th instance of variation to generate a list of 223 SNVs and 37 indels.

However, whilst all classes of SNVs were well represented, the diversity of indels has much lower representation due to the diversity and uniqueness of these events. Therefore, we secondly aimed to represent a wide range of variant-type diversity by additionally supplementing the list with 139 indels that were selected to provide a wider range of variant sizes. This enabled a large set of both SNVs and indels variants to be assessed using sequins. In total, 223 SNVs and 176 indels were represented.

The DNA sequence of selected human variants along with upstream and downstream flanking sequences were retrieved from the human genome sequence (Hg38)[7]. Sequences were then (i) inverted, (ii) shuffled in local windows and/or (iii) manually curated in order to abolish homology to their native genome positions (performed in order; see below).

Incorporating large structural variation (SVs). 12 examples of deletions ($n = 3$), insertions ($n = 3$), inversions ($n = 3$), mobile-element insertions ($n = 3$) and copy-number amplifications ($n = 3$) were randomly selected from the Database of Genomic Variants[27]. Each example was required to be less than 2 kb so that it could be fully encompassed within an individual sequin pair, and at least three examples of each different structural variant type were selected. The sequence of each SV, along with 1-kb flanking upstream and downstream sequences, were (i) inverted, (ii) shuffled in local windows and/or (iii) manually curated to abolish homology to known natural sequences (performed in order; see below). Notably, while performing inversion and/or shuffling, the integrity of any internal structure (such as repeat or inverted units) was maintained.

Design of mobile repeat elements. We retrieved natural human DNA sequences for five instances of mobile elements from common repeat classes (*AluSx*, *MIRb*, *L2a*, etc.) (http://repeatmasker.org). Repeat sequences were inverted and/or shuffled to remove homology (see below). Repeat elements were sufficiently sampled to achieve repeat density matching the human genome (Hg38). For example, an 8-Mb artificial chromosome sequence will have 788 *AluSx*, 534 *MIRb*, 433 *L2a*, 93 *MER5B* and 166 *L1M5* repeat mobile elements to match the density of analogous natural repeat elements.

Removing homology to natural sequences. To ensure that homology to natural genome sequences was successfully abolished following their inversion, artificial sequences were queried against the nucleotide collection database (nr/nt) using BLASTN[42] (word size = 28; expected threshold = 0.01; 1, −2, mismatch score and linear gap costs). Any sequences with significant matches to natural nucleotide sequences (expected value < 0.01) were subjected to local shuffling, nucleotide substitution and/or manual curation in order to abolish matches. Sequences with significant homology that could not be removed, for example, low-complexity sequences, were eventually omitted.

Assembly of in silico *chromosome.* Artificial sequences (background, small-scale genetic variation, structural variation and mobile repeat elements) were assembled into *in silico* chromosome sequences. Note that where possible, the distribution of features was matched to the distribution of corresponding natural examples in the human genome. Three *in silico* chromosome sequences were assembled; the first sequence (*chrIS_D*) contains no genetic variation and corresponds to the consensus reference sequence, analogous to the Hg38 reference sequence. This was used for alignment in subsequent analyses (see below). The remaining two sequences emulate diploid human genotypes, with homozygous variation encoded throughout both sequences and heterozygous variation encoded in only one of the sequences. The latter forms the template for synthetic DNA sequin synthesis (see below).

Generation of synthetic DNA sequins. The final sequence of each DNA sequin, along with their final concentrations in mixtures, can be accessed at http://www.sequin.xyz/. DNA sequin sequences are flanked by SapI Type II restriction digest sites. Each sequence was synthesized and inserted into a pMA vector by GeneArt (Life Technologies). Competent *E. coli* (Bioline) were thawed on ice and transformed with 2 µL of diluted NEBuilder HiFi DNA Assembly product per the manufacturer's suggested protocols. Transformed cells were plated on prewarmed 100 µg/mL ampicillin plates and incubated at 37 °C overnight (18 h). One colony from each plate was used to inoculate 5 mL LB broth containing 100 µg/mL ampicillin. Inoculated tubes were incubated overnight on a shaker at 37 °C. Plasmids were isolated using the Zyppy Plasmid Miniprep Kit (Zymo Research). The sequences of the purified plasmids were validated with Sanger sequencing and insert sequences excised by SapI Type II restriction digestion (New England BioLabs), according to manufacturer's protocols, and purified using Genomic DNA Clean & Concentrator kit (Zymo Research).

Preparation of DNA sequin mixtures. Purified DNA sequins were quantified using the BR dsDNA Qubit Assay on a Qubit 2.0 Fluorometer (Life Technologies) and verified on the Agilent 2100 Bioanalyzer with a Agilent High Sensitivity DNA Kit (Agilent Technologies). Individual DNA sequins were combined at differing concentrations using an epMotion 5070 liquid handling robot. Mixture stocks were prepared as single-use aliquots and stored at −80 °C.

Genomic DNA preparation. Cells of GM12878, an immortalized lymphoblast cell line derived from the NA12878 human individual, were purchased from Coriell Biorespotory (https://catalog.coriell.org/). GM12878 cells were cultured according to Coriell Cell Repositories growth protocols and standards. Briefly, GM12878 were cultured in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (FBS) at 37 °C

doi:10.1038/nmeth.3957

under 5% CO_2. DNA was extracted from GM12878 cells using TRIzol (Invitrogen) according to the manufacturer's instructions. Extracted DNA samples were treated with RNase A, followed by cleanup with Genomic DNA Clean & Concentrator kit (Zymo Research). Purified DNA was quantified using Nanodrop (Thermo Scientific).

Preparation and sequencing of DNA libraries with synthetic DNA variant sequins. DNA sequin mixtures were spiked into DNA sample according to specified concentration before library preparation. The Nextera XT Sample Prep Kit (Illumina) was used to prepare DNA libraries according to the manufacturer's instructions. Prepared libraries were quantified on Qubit (Invitrogen) and verified on Agilent 2100 Bioanalyzer with a Agilent High Sensitivity DNA Kit (Agilent Technologies). Libraries were sequenced on a HiSeq 2500 (Illumina) at the Kinghorn Centre for Clinical Genomics, Darlinghurst, New South Wales. Resulting .fastq files were trimmed using Trim Galore http://www.bioinformatics.babraham.ac.uk/projects/trim_galore (default parameters) before downstream analyses.

Generation of simulated sequencing libraries. Simulated Illumina paired-end libraries were prepared using the ART toolkit[10] (ChocolateCherryCake version; --paired --len 101 --mflen 350 --sdev 50 --seqSys HS25). Simulated libraries (i.e., those containing no genetic variant sites) were generated from sequin sequences (.fasta format) or from the Hg38 reference genome[7]. To generate simulated libraries containing human genetic variation, the human reference genome was amended with high-confidence SNVs and short indels from the annotated NA12878 genome using the GATK's FastaAlternateReferenceMaker tool before being used as a template for simulated library generation with ART (as above).

Experimental paired-end whole-genome sequencing libraries for the NA12878 individual were also obtained via the Illumina Platinum Genome Project (http://www.illumina.com/platinumgenomes; European Nucleotide Archive accession no. ERS179577).

Alignment of sequenced and simulated libraries. To coindex the artificial *in silico* chromosome with the human reference genome, *chrIS_D* was concatenated with Hg38 (.fasta format) before indexation using BWA index[11]. All alignment (sequenced and simulated libraries) was performed with the BWA mem algorithm[11] to combined or separate indices for *chrIS_D* or Hg38. Default parameters were used in all analyses, with the exception of those that assessed the potential for cross-alignment of sequin reads to Hg38 and human reads to *chrIS_D*. Here the alignment score penalty levied against mismatches (--B option) was incrementally reduced from 4 to 0 in order to force cross-alignment. Ambiguous alignments were those that were mapped with a MapQ score less than 60, which denotes the highest probability of correct alignment attainable using BWA at these parameters[11].

Assessment of sequenced reads quality. We used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (v0.10.1) to compare the per-base quality, per-sequence quality, and length distribution of reads emanating from sequins (aligned to *chrIS_D*) and NA12878 genome (aligned to GRCh38). Alignment quality rates were assessed according to MapQ scores[11].

Quantitative variability and optional correction. In order to investigate the magnitude of quantitative variation arising from either (i) sequence-specific biases (i.e., biases during library preparation, sequencing and alignment) or (ii) aliquot variability during the preparation of sequin mixtures by robot pipetting, two sequin flat mixtures were independently prepared. We assumed that sequence-specific variation impacts individual sequins in a reproducible manner in different mixtures, whereas aliquot variation is independent of standard identity and random between mixtures. The Bedtools coverage tool[43] was used to determine the median per-base coverage (which we take as the best measure of sequence abundance) of each standard within each mixture, with 100 nt excluded from each terminus to avoid sequencing edge effects[9]. The deviation of a standard from the median abundance within a mixture represents the sum of sequence-specific and sequence-independent variability. The average deviation for a given standard between the two mixtures represents the magnitude of its sequence-specific (systematic) quantitative bias, with the remaining variability in each mixture representing the contribution of aliquot variability.

Having calculated the magnitude of sequence-specific bias for each standard, it is possible to adjust observed abundances accordingly. To demonstrate this, reads aligning to each standard were randomly subsampled (Samtools –s option) according to their systematic bias correction score. Quantitative variability within each flat mixture was assessed following this correction, with overall variability being considerably reduced in each.

Coverage analysis and subsampling. Per-base coverage depth across the human genome and sequin sequences was determined using the Bedtools coverage tool[43]. Given the loss of coverage at sequin borders due to edge effects, we excluded 100 nt at each terminus from the analysis. Notably, no variants are located within terminal regions to omit the impact of edge effects[9]. Reads aligning to *chrIS_D* were randomly subsampled (samtools view -s option)[44] to achieve a median per-base coverage equivalent to the human NA12878 genome. To compare the coverage heterogeneity across sequin sequences with the human genome, we randomly selected 36× ~1,000 nt intervals roughly centered on annotated NA12878 variant sites within the Hg38 reference (comparable to 36× ~1,000 nt sequin standards) and generated population distributions of per-base coverage within these intervals. Median and interquartile range of per-base coverage was equivalent for sequins and their comparable intervals within the human genome.

Detecting homozygous and heterozygous variants. Variants were identified from alignments using the GATK[13–15]. In accordance with the GATKs 'best practice pipeline'[15], alignments were sorted, deduplicated, subjected to local realignment to improve indel detection and recalibrated before variant calling. Variant detection was performed in the absence of previous variant annotations (--genotyping_mode DISCOVERY). Putative SNVs and indels were filtered with the following parameters:

SNVs: --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < --12.5 || ReadPosRankSum < –8.0"

In/Dels: "QD < 2.0 || FS > 200.0 || ReadPosRankSum < –20.0"

doi:10.1038/nmeth.3957

Variants that passed filtering were divided into SNVs and indels and parsed according to genotype (homozygous or heterozygous) before being matched to known sites (GATK SelectVariants tool) of synthetic variation in *chrIS_D* or annotated high-confidence variants in the NA12878 genome (Illumina Platinum Genome Project). NA12878 variants falling outside the bounds of Platinum Genome high-confidence regions were not considered in this analysis. Sensitivity of variant detection was calculated as a percentage of known variant sites correctly retrieved. By progressively subsampling libraries (samtools view -s option)[44] and repeating the analysis as described above, we were able to assess the effect of depreciating sequencing coverage on variant detection sensitivity.

Detecting synthetic variants across allelic frequency range. Synthetic SNVs and small indels at heterogeneous allele frequencies were identified from alignments using the VarScan2 algorithm[22]. Alignments were converted into pileup format (samtools mpileup)[44] before searching for variants using VarScan. v2.3.9 (--min-coverage 5, --min-reads 1, --min-var-freq 0.00001). The VarScan2 output contains all possible variants, including false-positive variants erroneously identified from sequencing and alignment errors. These were filtered according to a range of possible criteria, including supporting read counts, allele frequency, the ratio of supporting reads on each strand, the quality of variant-supporting nucleotide calls and confidence thresholds (*P* values), which are calculated by VarScan2 according to a Fisher's exact test[22].

We ranked true- and false-positive variants according to confidence threshold (*P* values) to determine true-positive and false-positive discovery rates as a function of minimum-confidence threshold (*P* values), either for the total SNV and indel pools or within groups parsed according to expected allelic frequencies. To determine the sensitivity and precision of variant detection within each allelic fraction, we generated receiver operating characteristic (ROC) and limit of detection ratio (LODR) plots using ROCR[45] and R ggplot2.

Structural variation. Total alignments were parsed into split reads (extractSplitReads_BwaMem tool in the Lumpy package[28]), discordant read pairs (samtools filtering flag: -F 1294)[44] and 'hanging reads', in which one pair end (but not its mate) was uniquely aligned (samtools filtering flags –f 8 –F 260)[44]. Different alignment types were visualized as a wiggle track with the Integrated Genome Viewer (IGV) at annotated breakpoints in the *in silico* chromosome and NA12878 genome (for insertions and deletions). NA12878 annotations were downloaded from Genome In A Bottle (https://github.com/genome-in-a-bottle/giab_data_indexes).

We used the Lumpy package[28] to identify deletions, inversions and copy-number amplifications with respect to *chrIS_D* with the following parameters:

Discordant pairs: read_length:125,min_non_overlap:125, discordant_z:5,back_distance:10,min_mapping_threshold:20, weight:1

Split reads: back_distance:10,min_mapping_threshold:20, weight:1

Because insertions are too large to accrue spanning split or discordant read-pairs, they were instead characterized by the accumulation of hanging reads, which converge from either direction onto the insertion site.

Copy-number variants and quantitative DNA scale. Total alignments from *chrIS_D* we parsed into split reads (extractSplitReads_BwaMem tool in the Lumpy package[28]) and discordant read pairs (samtools filtering flag: –F 1294)[44].

Annotated structural variants within the NA12878 genome were retrieved from Genome In A Bottle (https://github.com/genome-in-a-bottle/giab_data_indexes) and Lumpy (**Supplementary Data 5** 13059_2013_3363_MOESM5_ESM) made with Lumpy[44] for NA12878 with paired-end and split-read evidence that were also validated with PacBio/Moleculo data.

We applied the mean-shift algorithm implemented in CNVnator[46] to assess significant mean-coverage fold change with respect to the genome background. The algorithm compares the coverage for each nucleotide within a specified region, as well as mean coverage and s.d. across the region specified to background average using a one-sample *t*-test, adjusted for multiple testing. The background regions can comprise randomly selected, size-matched genome control regions (that do not overlap RepeatMasker annotations) or a query set of CNV sequins.

Limit of quantification was determined using piece-wise linear regression test. DNA repeat sequences were retrieved from GenBank using the following accession IDs: rDNA (U13369), chrM (J01415.2), chrEBV (V01555.2); MSR5p (AC106774.2), FLJ40296 (AL353652.17), DXZ4 (S60754.1), CT47 (AL670379.17).

Statistical analysis and graph plotting. GraphPad Prism (GraphPad Software, La Jolla, California, USA) (v6.0f) and R (v3.1.3), including kernel density methods available in ggplot2, were used to generate plots and perform statistical calculations presented in figures and main text.

Code availability. All associated data files, including *in silico* chromosome and sequin sequences, variant annotations and sequencing libraries can be accessed at http://www.sequin.xyz/.

42. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
43. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
46. 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).

doi:10.1038/nmeth.3957