(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
06 July 2023 (06.07.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/129953 A2

(51) International Patent Classification:
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)

(21) International Application Number:
PCT/US2022/082462

(22) International Filing Date:
28 December 2022 (28.12.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

63/294,813	29 December 2021 (29.12.2021)	US
63/294,816	29 December 2021 (29.12.2021)	US
63/294,820	29 December 2021 (29.12.2021)	US
63/294,827	29 December 2021 (29.12.2021)	US
63/294,828	29 December 2021 (29.12.2021)	US
63/294,830	29 December 2021 (29.12.2021)	US
17/952,192	23 September 2022 (23.09.2022)	US
17/952,194	23 September 2022 (23.09.2022)	US
17/952,198	23 September 2022 (23.09.2022)	US

(71) Applicants: **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US). **ILLUMINA CAMBRIDGE LIMITED** [GB/GB]; 19 Granta Park, Great Abington, Cambridge CB21 6DF (GB).

(72) Inventors: **GAO, Hong**; 5200 Illumina Way, San Diego, California 92122 (US). **HAMP, Tobias**; 19 Granta Park, Great Abington, Cambridge CB21 6DF (GB). **MCMASTER-SCHRAIBER, Joshua Goodwin Jon**; 5200 Illumina Way, San Diego, California 92122 (US). **SUNDARAM, Laksshman**; 5200 Illumina Way, San Diego, California 92122 (US). **FARH, Kai-How**; 5200 Illumina Way, San Diego, California 92122 (US).

(74) Agent: **FUNK, Christopher** et al.; Keller Preece PLLC, 1010 N 500 E, STE 210, North Salt Lake, Utah 84054 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

(54) Title: VARIANT CALLING WITHOUT A TARGET REFERENCE GENOME

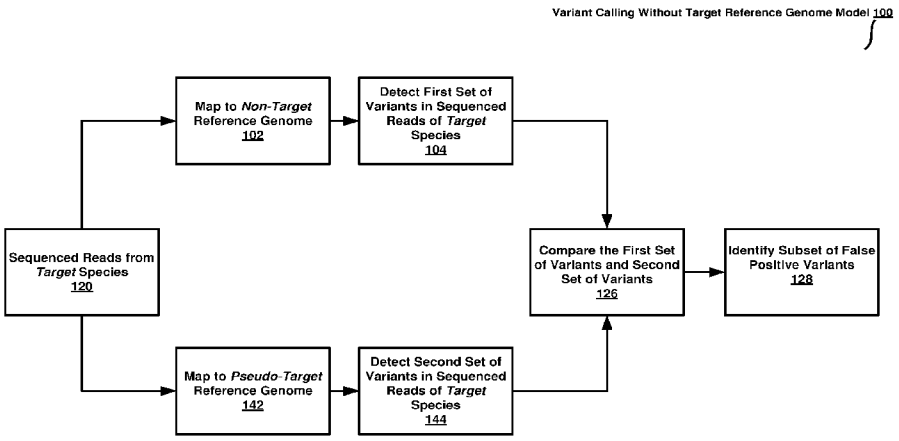


Figure 1

(57) **Abstract:** The technology disclosed relates to determining feasibility of using a reference genome of a non-target species for variant calling a sample of a target species. In particular, the technology disclosed relates to mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species, and mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species.

WO 2023/129953 A2

[Continued on next page]

HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

VARIANT CALLING WITHOUT A TARGET REFERENCE GENOME

PRIORITY APPLICATIONS

[0001] This application claims the benefit of and priority to the following:

[0002] U.S. Patent Application No. 17/952,192, titled, "VARIANT CALLING WITHOUT A TARGET REFERENCE GENOME," filed September 23, 2022 (Attorney Docket No. ILLM 1064-2/IP-2297-US1);

[0003] U.S. Patent Application No. 17/952,194, titled, "QUALITY DETECTION OF VARIANT CALLING USING A MACHINE LEARNING CLASSIFIER," filed September 23, 2022 (Attorney Docket No. ILLM 1064-3/IP-2297-US2);

[0004] U.S. Patent Application No. 17/952,198, titled, "UNIQUE MAPPER TOOL FOR EXCLUDING REGIONS WITHOUT ONE-TO-ONE MAPPING BETWEEN A SET OF TWO REFERENCE GENOMES," filed September 23, 2022 (Attorney Docket No. ILLM 1064-4/IP-2297-US3);

[0005] U.S. Provisional Patent Application No. 63/294,813, titled "PERIODIC MASK PATTERN FOR REVELATION LANGUAGE MODELS," filed December 29, 2021 (Attorney Docket No. ILLM 1063-1/IP-2296-PRV);

[0006] U.S. Provisional Patent Application No. 63/294,816, titled "CLASSIFYING MILLIONS OF VARIANTS OF UNCERTAIN SIGNIFICANCE USING PRIMATE SEQUENCING AND DEEP LEARNING," filed December 29, 2021 (Attorney Docket No. ILLM 1064-1/IP-2297-PRV);

[0007] U.S. Provisional Patent Application No. 63/294,820, titled "IDENTIFYING GENES WITH DIFFERENTIAL SELECTIVE CONSTRAINT BETWEEN HUMANS AND NON-HUMAN PRIMATES," filed December 29, 2021 (Attorney Docket No. ILLM 1065-1/IP-2298-PRV);

[0008] U.S. Provisional Patent Application No. 63/294,827, titled "DEEP LEARNING NETWORK FOR EVOLUTIONARY CONSERVATION," filed December 29, 2021 (Attorney Docket No. ILLM 1066-1/IP-2299-PRV);

[0009] U.S. Provisional Patent Application No. 63/294,828, titled "INTER-MODEL PREDICTION SCORE RECALIBRATION," filed December 29, 2021 (Attorney Docket No. ILLM 1067-1/IP-2301-PRV); and

[0010] U.S. Provisional Patent Application No. 63/294,830, titled "SPECIES-DIFFERENTIABLE EVOLUTIONARY PROFILES," filed December 29, 2021 (Attorney Docket No. ILLM 1068-1/IP-2302-PRV).

[0011] The priority applications are incorporated by reference as if fully set forth herein.

FIELD OF THE TECHNOLOGY DISCLOSED

[0012] The technology disclosed relates to artificial intelligence type computers and digital data processing systems and corresponding data processing methods and products for emulation of intelligence (*i.e.*, knowledge based systems, reasoning systems, and knowledge acquisition systems); and including systems for reasoning with uncertainty (*e.g.*, fuzzy logic systems), adaptive systems, machine learning systems, and artificial neural networks. In particular, the technology disclosed relates to using deep convolutional neural networks to analyze ordered data.

INCORPORATIONS

[0013] The following are incorporated by reference for all purposes as if fully set forth herein, and should be considered part of, this provisional patent filing:

[0014] Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161–1170 (2018);

[0015] Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535–548 (2019);

[0016] US Patent Application No. 62/573,144, titled "TRAINING A DEEP PATHOGENICITY CLASSIFIER USING LARGE-SCALE BENIGN TRAINING DATA," filed October 16, 2017 (Attorney Docket No. ILLM 1000-1/IP-1611-PRV);

[0017] US Patent Application No. 62/573,149, titled "PATHOGENICITY CLASSIFIER BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)," filed October 16, 2017 (Attorney Docket No. ILLM 1000-2/IP-1612-PRV);

[0018] US Patent Application No. 62/573,153, titled "DEEP SEMI-SUPERVISED LEARNING THAT GENERATES LARGE-SCALE PATHOGENIC TRAINING DATA," filed October 16, 2017 (Attorney Docket No. ILLM 1000-3/IP-1613-PRV);

[0019] US Patent Application No. 62/582,898, titled "PATHOGENICITY CLASSIFICATION OF GENOMIC DATA USING DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)," filed November 7, 2017 (Attorney Docket No. ILLM 1000-4/IP-1618-PRV);

[0020] US Patent Application No. 16/160,903, titled "DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS," filed on October 15, 2018 (Attorney Docket No. ILLM 1000-5/IP-1611-US);

[0021] US Patent Application No. 16/160,986, titled "DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION," filed on October 15, 2018 (Attorney Docket No. ILLM 1000-6/IP-1612-US);

[0022] US Patent Application No. 16/160,968, titled "SEMI-SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS," filed on October 15, 2018 (Attorney Docket No. ILLM 1000-7/IP-1613-US);

[0023] US Patent Application No. 16/160,978, titled "DEEP LEARNING-BASED SPLICE SITE CLASSIFICATION," filed on October 15, 2018 (Attorney Docket No. ILLM 1001-4/IP-1680-US);

[0024] US Patent Application No. 16/407,149, titled "DEEP LEARNING-BASED TECHNIQUES FOR PRE-TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS," filed May 8, 2019 (Attorney Docket No. ILLM 1010-1/IP-1734-US);

[0025] US Patent Application No. 17/232,056, titled "DEEP CONVOLUTIONAL NEURAL NETWORKS TO PREDICT VARIANT PATHOGENICITY USING THREE-DIMENSIONAL (3D) PROTEIN STRUCTURES," filed on April 15, 2021, (Atty. Docket No. ILLM 1037-2/IP-2051-US);

[0026] US Patent Application No. 63/175,495, titled "MULTI-CHANNEL PROTEIN VOXELIZATION TO PREDICT VARIANT PATHOGENICITY USING DEEP CONVOLUTIONAL NEURAL NETWORKS," filed on April 15, 2021, (Atty. Docket No. ILLM 1047-1/IP-2142-PRV);

[0027] US Patent Application No. 63/175,767, titled "EFFICIENT VOXELIZATION FOR DEEP LEARNING," filed on April 16, 2021, (Atty. Docket No. ILLM 1048-1/IP-2143-PRV);

[0028] US Patent Application No. 17/468,411, titled "ARTIFICIAL INTELLIGENCE-BASED ANALYSIS OF PROTEIN THREE-DIMENSIONAL (3D) STRUCTURES," filed on September 7, 2021, (Atty. Docket No. ILLM 1037-3/IP-2051A-US);

[0029] U.S. Provisional Patent Application No.: 63/253,122, titled "PROTEIN STRUCTURE-BASED PROTEIN LANGUAGE MODELS," filed October 6, 2021 (Attorney Docket No. ILLM 1050-1/IP-2164-PRV);

[0030] U.S. Provisional Patent Application No.: 63/281,579, titled "PREDICTING VARIANT PATHOGENICITY FROM EVOLUTIONARY CONSERVATION USING THREE-DIMENSIONAL (3D) PROTEIN STRUCTURE VOXELS," filed November 19, 2021 (Attorney Docket No. ILLM 1060-1/IP-2270-PRV); and

[0031] U.S. Provisional Patent Application No.: 63/281,592, titled "COMBINED AND TRANSFER LEARNING OF A VARIANT PATHOGENICITY PREDICTOR USING GAPED AND NON-GAPED PROTEIN SAMPLES," filed November 19, 2021 (Attorney Docket No. ILLM 1061-1/IP-2271-PRV).

BACKGROUND

[0032] The subject matter discussed in this section should not be assumed to be prior art merely as a result of its mention in this section. Similarly, a problem mentioned in this section or associated with the subject matter provided as background should not be assumed to have been previously recognized in the prior art. The subject matter in this section merely represents different approaches, which in and of themselves can also correspond to implementations of the claimed technology.

[0033] The explosion of available biological sequence data has led to multiple computational approaches that infer the proteins' three-dimensional structure, biological function, fitness, and evolutionary history from sequence data. So-called protein language models, like the ones based on the Transformer architecture, have been trained on large ensembles of protein sequences by using the masked language modeling objective of filling in masked amino acids in a sequence, given the surrounding ones.

[0034] Protein language models capture long-range dependencies, learn rich representations of protein sequences, and can be employed for multiple tasks. For example, protein language models can predict structural contacts from single sequences in an unsupervised way.

[0035] Protein sequences can be classified into families of homologous proteins that descend from an ancestral protein and share a similar structure and function. Analyzing multiple sequence alignments (MSAs) of homologous proteins provides important information about functional and structural constraints. The statistics of MSA columns, which represent amino-acid sites, identify functional residues that are conserved during evolution. Correlations of amino acid usage between the MSA columns contain important information about functional sectors and structural contacts.

[0036] Language models were initially developed for natural language processing and operate on a simple but powerful principle: they acquire linguistic understanding by learning to fill in missing words in a sentence, akin to a sentence completion task in standardized tests. Language models develop powerful reasoning capabilities by applying this principle across large text corpora. The Bidirectional Encoder Representations from Transformers (BERT) mode instantiated this principle using Transformers, a class of neural networks in which attention is the primary component of the learning system. In a Transformer, each token in the input sentence can "attend" to all other tokens by exchanging activation patterns corresponding to the intermediate outputs of neurons in a neural network.

[0037] Protein language models like the MSA Transformer have been trained to perform inference from MSAs of evolutionarily related sequences. The MSA Transformer interleaves per-sequence ("row") attention with per-site ("column") attention to incorporate coevolution.

Combinations of row attention heads in the MSA Transformer have led to state-of-the-art unsupervised structural contact predictions.

[0038] End-to-end deep learning approaches for variant effect predictions are applied to predict the pathogenicity of missense variants from protein sequence and sequence conservation data (*See* Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161–1170 (2018), referred to herein as "PrimateAI"). PrimateAI uses deep neural networks trained on variants of known pathogenicity with data augmentation using cross-species information. PrimateAI in particular uses sequences of wild-type and mutant proteins to compare the difference and decide the pathogenicity of mutations using the trained deep neural networks. Such an approach that utilizes the protein sequences for pathogenicity prediction is promising because it can avoid the problem of circularity and overfitting to previous knowledge. However, compared to the adequate number of data to train the deep neural networks effectively, the number of clinical data available in ClinVar is relatively small. To overcome this data scarcity, PrimateAI uses common human variants and variants from primates as benign data while simulated variants based on trinucleotide context were used as unlabeled data.

[0039] PrimateAI outperforms prior methods when trained directly upon sequence alignments. PrimateAI learns important protein domains, conserved amino acid positions, and sequence dependencies directly from the training data consisting of about 120,000 human samples. PrimateAI substantially exceeds the performance of other variant pathogenicity prediction tools in differentiating benign and pathogenic de-novo mutations in candidate developmental disorder genes, and in reproducing prior knowledge in ClinVar. These results suggest that PrimateAI is an important step forward for variant classification tools that may lessen the reliance of clinical reporting on prior knowledge.

[0040] Therefore, an opportunity arises to use protein language models and MSAs for variant pathogenicity prediction. More accurate variant pathogenicity prediction may result.

BRIEF DESCRIPTION OF THE DRAWINGS

[0041] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee. The color drawings also may be available in PAIR via the Supplemental Content tab.

[0042] In the drawings, like reference characters generally refer to like parts throughout the different views. Also, the drawings are not necessarily to scale, with an emphasis instead generally being placed upon illustrating the principles of the technology disclosed. In the following description, various implementations of the technology disclosed are described with reference to the following drawings, in which.

[0043] Figure 1 is a flow diagram that illustrates a process of a system for variant calling for a particular Target Species 120 without using the target reference genome.

[0044] Figure 2 is a sequential flow diagram representing the process of identifying false positive variants by means of a set comparison for identification of true positive variants from sequenced reads.

[0045] Figure 3 illustrates an example of genetic variants from an example reference genetic sequence A with two example variant sequences in which the variant sequences possess a respective single nucleotide variant in a single base position but otherwise possess an identical composition to the reference sequence.

[0046] Figure 4 graphically represents the process of base-resolution variant calling sensitivity detection in a series of example flow diagrams mapping sequenced reads to reference genomes.

[0047] Figure 5 is a schematic illustrating the process of mapping sequenced reads to a reference genome.

[0048] Figure 6 represents a graphical flow diagram of a process for alternative detection of the second variant set in one implementation of the technology disclosed where a target species reference genome is available.

[0049] Figure 7 is a schematic illustrating the evolutionary relationship between the target species and the pseudo-target species using simplified phylogenic tree graphics.

[0050] Figure 8 is a schematic illustrating the evolutionary relationship between the target species and the pseudo-target species using a simplified phylogenic tree graphic.

[0051] Figure 9 is a schematic representing genome-resolution variant calling sensitivity detection of sequenced reads.

[0052] Figure 10 is a schematic flow diagram demonstrating how the output data generated by the set comparison of sequenced reads mapped to a non-target reference genome and a pseudo-target reference genome can be used as a training dataset for a quality classifier to predict variant quality where a reference genome is not available, in contrast to methods wherein the quality is determined by mapping to non-target or pseudo-target genomes.

[0053] Figure 11 is an illustrative example of the variant features in the plurality of variant features describing guanine-cytosine content.

[0054] Figure 12 is an illustrative example of the variant feature in the plurality of variant features describing local composition complexity.

[0055] Figure 13 is an illustrative example of the variant feature in the plurality of variant features describing allelic count.

[0056] Figure 14 is an illustrative example of the variant feature in the plurality of variant features describing process of mapping sequenced reads to a reference genome wherein an additional step is added to compute a quality metric of the mapping.

[0057] Figure 15 is an illustrative example of the variant feature in the plurality of variant features describing strand bias in sequenced reads when mapped to a reference genome.

[0058] Figure 16 is an illustrative example of the variant features in the plurality of variant features describing the depth and coverage of sequenced reads mapped to a reference genome.

[0059] Figure 17 is an illustrated flow diagram representing the variant quality classifier configured as a random forest model to classify a target variant as either belonging to the high quality class or the low quality class.

[0060] Figure 18 is an illustrated flow diagram representing the variant quality classifier configured as a logistic regression model to classify a target variant as either belonging to the high quality class or the low quality class.

[0061] Figure 19 is an illustrated flow diagram representing the variant quality classifier configured as a neural network model to classify a target variant as either belonging to the high quality class or the low quality class.

[0062] Figure 20 is a flow diagram of the unique mapper overview process to further improve the quality of a set of called variants following variant calling or machine learning variant classification.

[0063] Figure 21 schematically illustrates the gene annotation filter in the series of cascaded filters for variant quality filtering.

[0064] Figure 22 schematically illustrates the process codon transcription and translation and filtering for codon match.

[0065] Figure 23 illustrates the process of filtering genes based on a distribution of machine learning scores.

[0066] Figure 24 illustrates deviation from Hardy-Weinberg Equilibrium in an example population.

[0067] Figure 25 is an illustration of a nonsense variant.

[0068] Figure 26 contains a graph of collected results demonstrating the cascading filter effect on the number of nonsense variants per sample.

[0069] Figure 27 contains a graph of collected results demonstrating the cascading filter effect on missense: synonymous ratio of called variants per sample.

[0070] Figure 28 contains a graph of collected results demonstrating the cascading filter effect on number of insertion-deletion variants (indels) per sample.

[0071] Figure 29 shows an example computer system that can be used to implement the technology disclosed.

DETAILED DESCRIPTION

[0072] The following discussion is presented to enable any person skilled in the art to make and use the technology disclosed and is provided in the context of a particular application and its requirements. Various modifications to the disclosed implementations will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other implementations and applications without departing from the spirit and scope of the technology disclosed. Thus, the technology disclosed is not intended to be limited to the implementations shown but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0073] The detailed description of various implementations will be better understood when read in conjunction with the appended drawings. To the extent that the figures illustrate diagrams of the functional blocks of the various implementations, the functional blocks are not necessarily indicative of the division between hardware circuitry. Thus, for example, one or more of the functional blocks (*e.g.*, modules, processors, or memories) may be implemented in a single piece of hardware (*e.g.*, a general purpose signal processor or a block of random access memory, hard disk, or the like) or multiple pieces of hardware. Similarly, the programs may be stand-alone programs, may be incorporated as subroutines in an operating system, may be functions in an installed software package, and the like. It should be understood that the various implementations are not limited to the arrangements and instrumentality shown in the drawings.

[0074] The processing engines and databases of the figures, designated as modules, can be implemented in hardware or software, and need not be divided up in precisely the same blocks as shown in the figures. Some of the modules can also be implemented on different processors, computers, or servers, or spread among a number of different processors, computers, or servers. In addition, it will be appreciated that some of the modules can be combined, operated in parallel or in a different sequence than that shown in the figures without affecting the functions achieved. The modules in the figures can also be thought of as flowchart steps in a method. A module also need not necessarily have all its code disposed contiguously in memory; some parts of the code can be separated from other parts of the code with code from other modules or other functions disposed in between.

[0075] The technologies disclosed can be used to improve the quality of pathogenic variant calling. The technology disclosed can be used to improve the quality of variant calling in scenarios where desired reference genomes are unavailable. There are 8.7 million species worldwide, but very few have reference genome builds. In many scenarios, we need to do variant

calling in the absence of reference genome builds. In some instances, we could choose a closely-related species as a reference genome for variant calling. But this, sometimes, leads to many false positive calls. Thus, we developed various methods to reduce the false positives, including the random forest classifiers, linear regression models, and neural network models. We also devised a unique-mapper score to identify regions that are not one-to-one mapping between the species, which will further reduce variant calling errors.

[0076] In some embodiments, the technology disclosed relates to determining feasibility of using a reference genome of a non-target species for variant calling a sample of a target species. In particular, the technology disclosed relates to mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species, and mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species.

[0077] Further, in certain embodiments, the technology disclosed relates to variant calling of sequenced reads of a sample of a target species against a reference genome of a pseudo-target species. Low-quality variants are identified as false positive variants that are present in the second set of variants but absent from the first set of variants.

[0078] Additionally, in some cases, the technology disclosed relates to a variant quality classifier that is configured to process a plurality of features of a target variant and generate a quality indication for the quality variant. The variant quality classifier is trained on a set of high-quality variants and a set of low-quality variants. High-quality variants are identified as true positive variants that are common between a first set of variants detected by variant calling sequenced reads of a sample of a target species against a reference genome of a non-target species and a second set of variants detected by variant calling of sequenced reads of a sample of a target species against a reference genome of a pseudo-target species. Low-quality variants are identified as false positive variants that are present in the second set of variants but absent from the first set of variants.

Variant Calling Using a Non-Target Reference Genome

[0079] Figure 1 is a flow diagram that illustrates a process 100 of a system for variant calling for a particular Target Species 120 without using the target reference genome. Mapping sequenced reads from a Target Species 120 to a Non-Target Reference Genome 102 detects a First Set of Variants in the Sequenced Reads of the Target Species 104. The Non-Target Reference Genome 102 is from a non-target species other than the Target Species 120. In some implementations of the technology disclosed, the Non-Target Reference Genome 102 is non-homologous with the genome of Target Species 120, as determined by a homology threshold

(such as a percentage homology below 30%, 40%, or 50%, or a double-bounded range of acceptable homology percentages such as 30-40% or 40-50%). In certain embodiments, the non-target species and Target Species 120 belong to the same taxonomical genus, family, order, or class.

[0080] Mapping sequenced reads from a Target Species 120 to a Pseudo-Target Reference Genome 142 detects a Second Set of Variants in the Sequenced Reads of the Target Species 144. The Pseudo-Target Reference Genome 142 is from a pseudo-target species other than the Target Species 120. In some implementations of the technology disclosed, the Pseudo-Target Reference Genome 142 is homologous with the genome of Target Species 120, as determined by a homology threshold (such as a percentage homology above 80%, 90%, or 95%, or a double-bounded range of acceptable homology percentages such as 85-90% or 80-89%). A homology threshold set to determine degree of homology between the pseudo-target species and target species may be the same as a homology threshold set to determine degree of homology between the non-target species and target species, or the respective homology thresholds may differ. In some embodiments, the homology threshold set to determine degree of homology between the non-target species and target species may be informed by the degree of homology between the pseudo-target species and target species, or vice versa. The Comparison 126 of the first set of variants and second set of variants identifies a subset of False Positive Variants 128 (*i.e.*, overlapping variants identified by mapping to the Pseudo-Target Reference Genome 142 cannot be considered as reliable positive variants on the basis of homology when the variants are also identified by mapping to Non-Target Reference Genome 102).

[0081] Figure 2 is a sequential flow diagram 200 representing the process of identifying false positive variants by means of a set comparison for identification of true positive variants from sequenced reads. Mapping sequenced reads from a Target Species 202 to a Non-Target Species Reference Genome 204 detects a First Set of Variants 223. Mapping sequenced reads from a Target Species 206 to a Pseudo-Target Species Reference Genome 208 detects a Second Set of Variants 227. A Venn diagram represents the union Variant Set 1 ∪ Variant Set 2, wherein the set difference Variant Set 1 – Variant Set 2 is represented by the left downward diagonal shaded area 244 and the set difference Variant Set 2 – Variant Set 1 is represented by the right upward diagonal shaded area 246. The intersection Variant Set 1 ∩ Variant Set 2 is represented by the center diamond crosshatch shaded area for a set of True Positive Variants 266. The intersection Variant Set 1 ∩ Variant Set 2 (i.e., called variants identified both in the first set of variants 223 detected by mapping to a Non-Target Reference Genome 204 and the Second Set of Variants 227 detected by mapping to a Pseudo-Target Reference Genome 208) translates to the set of True Positive Variants 266. The set difference Variant Set 2 – Variant Set 1 (i.e., called variants

identified in the Second Set of Variants 227 detected by mapping to a Pseudo-Target Reference Genome 208 but not identified in the First Set of Variants 223 detected by mapping to a Non-Target Reference Genome 204) translates to the Set of False Positive Variants 268.

[0082] Figure 3 illustrates an example of Genetic Variants 300 from an example Reference Genetic Sequence A 302 with two example Variant Sequences 1A 322 and 1B 342 in which the variant sequences possess a respective single nucleotide variant in a single base position but otherwise possess an identical composition to the reference sequence. For example, a single nucleotide substitution is shown as Adenine 326 in Variant 1A 322 and Thymine 336 in Variant 1B 342 as compared to Cytosine 306 in the Reference Genetic Sequence A 302.

[0083] Figure 4 graphically represents the process 400 of base-resolution variant calling sensitivity detection in a series of example flow diagrams mapping sequenced reads to reference genomes. Sequenced Read X 410 is mapped to Pseudo-Target Reference Genome 402 and will be called as a variant due to the A→C single nucleotide variant in position 5. Sequenced Read X 410 is also mapped to Non-Target Reference Genome 404 and will not be called as a variant as a single nucleotide variant is not identified. As a result, Sequenced Read X 410 belongs to the set difference between the called variant set from mapping to the Pseudo-Target Reference Genome 402 and the variant set from mapping to the Non-Target Reference Genome 404 therefore Sequenced Read X 410 is a false positive. Sequenced Read Y 412 is mapped to Pseudo-Target Reference Genome 422 and will be called as a variant due to the A→C single nucleotide variant in position 5. Sequenced Read Y 412 is also mapped to Non-Target Reference Genome 424 and will be called as a variant due to the A→C single nucleotide variant in position 5. As a result, Sequenced Read Y 412 belongs to the set intersection between the called variant set from mapping to the Pseudo-Target Reference Genome 422 and the variant set from mapping to the Non-Target Reference Genome 424 therefore Sequenced Read Y 412 is a true positive.

[0084] Sequenced Read Z is mapped to Pseudo-Target Reference Genome 442 and will not be called as a variant despite the cytosine and guanine not being equivalent at position five. Due to base pairing, the complementary strand of the Pseudo-Target Reference Genome 442 possesses a cytosine at position 5 and the complementary strand of the Sequenced Read Z 414 possesses a guanine at position 5. As a result, this Sequenced Read Z 414 is not a variant when mapped to Pseudo-Target Reference Genome 442. Sequenced Read Z 414 is also mapped to Non-Target Reference Genome 444 and will not be called as a variant due to complementary bases being present at position 5. As a result, Sequenced Read Z 414 belongs to the complement of both the called variant set from mapping to the Pseudo-Target Reference Genome 442 and the called variant set from the Non-Target Reference Genome 444 therefore Sequenced Read Z 414 is a true negative variant.

[0085] Figure 5 is a schematic 500 illustrating the process of mapping sequenced reads to a Reference Genome. Sequenced Reads 502 are mapped to Reference Genome 505 resulting in Mapping 555. In Mapping 555, each sequenced read from the set of Sequenced Reads 502 aligns with a given genomic region within the Reference Genome 505. As seen in Mapping 555, mapped sequenced reads may align with genomic regions that are mutually exclusive (*i.e.*, do not overlap, such as the leftmost sequenced read and the middle sequenced read in Mapping 555) or are not mutually exclusive from each other (*i.e.*, do overlap, such as the middle sequenced read and the rightmost sequenced read in Mapping 555).

[0086] Figure 6 represents a graphical flow diagram 600 of a process for alternative detection of Variant Set Two in one implementation of the technology disclosed where a Target Species Reference Genome 622 is available. Sequenced Reads from A Target Species 602 are mapped to the Target Species Reference Genome 622. The mapped sequence reads of the target species are lifted over to a Non-Target Species Reference Genome 642. Variants that are detected in the Non-Target Species Reference Genome 642 but not detected in the Target Species Reference Genome 622 comprise Variant Set Two 662 wherein the plurality of variants in Variant Set Two 662 are false positive variants. The alternative detection of Variant Set Two 662 is useful in generating training data comprising known ground truth data to be used in training a machine learning classifier for the detection of true positive variants and false positive variants.

[0087] Figure 7 is a schematic 700 illustrating the evolutionary relationship between the target species and the pseudo-target species using simplified phylogenic tree graphics. In Phylogenic Tree A 702, the pseudo-target species and the target species are different species that are not orthologous. In one implementation of the technology disclosed, the evolutionary relationship between the target species and the pseudo-target species is reflective of that shown in Phylogenic Tree A 702. In Phylogenic Tree B 704, the pseudo-target species and the target species are different species that are orthologous. In one implementation of the technology disclosed, the evolutionary relationship between the target species and the pseudo-target species is reflective of that shown in Phylogenic Tree B 704. In Phylogenic Tree C 706, the pseudo-target species and the target species are the same species. In one implementation of the technology disclosed, the evolutionary relationship between the target species and the pseudo-target species is reflective of that shown in Phylogenic Tree C 706.

[0088] Figure 8 is a schematic 800 illustrating the evolutionary relationship between the target species and the pseudo-target species using a simplified phylogenic tree graphic. In Phylogenic Tree D 802, the non-target species and the target species are different species wherein the non-target species is a human and the target species is a non-human primate. In one implementation of the technology disclosed, the evolutionary relationship between the target species and the non-

target species reflects that of Phylogenic Tree D 802. In one implementation of the technology disclosed, primate species samples and reference genomes are leveraged to infer the pathogenicity of orthologous human variants by variant calling to closely-related primate species genomes, variant calling to non-target, non-homologous primate species genomes, and contrasting the results as demonstrated within Figures 1-5. In some implementations of the technology disclosed, machine learning classifiers are trained to detect false positive variants, further refining the identification process of true positive variants.

[0089] Figure 9 is a schematic 900 representing genome-resolution variant calling sensitivity detection of sequenced reads. Sequenced Read A 902 from a Target Species is equivalent to Sequenced Read A 942 from a Target Species and maps to Region One 922 in the Non-Target Reference Genome 924 and Region Two 962 in the Pseudo-Target Reference Genome 964. Region One 922 and Region Two 962 are not equivalent (i.e., not orthologous), therefore Sequenced Read A does not map to the same genomic region in the Non-Target Reference Genome 924 and in the Pseudo-Target Reference Genome 964. Despite mapping to both the Non-Target Reference Genome 924 and the Pseudo-Target Reference Genome 964, Sequenced Read A 902 does not result in a called variant in the genomic region within the Non-Target Reference Genome 964 that is orthologous to the genomic region that Sequenced Read A 942 maps to within the Pseudo-Target Reference Genome 964. As a result, this variant will belong to the called variant set from mapping to the Pseudo-Target Reference Genome 964 but will not belong to the called variant set from mapping to the Non-Target Reference Genome 924 and results in a false positive.

[0090] Sequenced Read B 982 from a Target Species, Sequenced Read B from a Target Species 984, and Sequenced Read B from a Target Species 986 are equivalent. Region Three 972, Region Four 978, and Region Five 980 belong to the non-target reference genome and are not equivalent. Sequenced Read B 982 from the Target Species maps to multiple regions within the non-target reference genome. As with Sequenced Read A 902, Sequenced Read B 982 will map to a different genomic region within the non-target species reference genome than the orthologous genomic region that Sequenced Read B 982 maps to within the pseudo-target reference genome due to the multiplicity of variant calling within the non-target reference genome. Subsequently, sequenced read that maps to more than two genomic regions within the non-target reference genome will result in a false positive.

Machine Learning Classifiers

[0091] Figure 10 is a schematic flow diagram 1000 demonstrating how the output data generated by the set comparison of sequenced reads mapped to a non-target reference genome and a pseudo-target reference genome can be used as a training dataset for a quality classifier to

predict variant quality where a reference genome is not available, in contrast to methods wherein the quality is determined by mapping to non-target or pseudo-target genomes. Sequenced Reads from a Target Species 1002 are mapped to reference genomes as previously described in Figures 1, 2, and 3. The intersection of Variant Set One 1004 and Variant Set Two 1006 corresponds to the Set of True Positive Variants 1022. The set difference between Variant Set Two 1006 and Variant Set One 1004 (i.e., present in Variant Set Two 1006 but not present in Variant Set One 1004) corresponds to the Set of False Positive Variants 1008. The Set of True Positive Variants 1022 is further coded as a Set of High Quality Variants 1024. The Set of False Positive Variants 1008 is further coded as a Set of Low Quality Variants 1010. The combined set of High Quality Variants 1024 and Low Quality Variants 1010 comprise the set of Ground Truth Data 1020.

[0092] The Quality Classifier 1064 undergoes a Model Training Process 1040 on the Ground Truth Data 1020. The Quality Classifier 1064 takes an Input Target Variant 1062 represented as a vector containing the set of variant features in the plurality of variant features $\{x_1:x_n\}$ where each value of x is a variant feature within the set of variant features in the plurality of variant features describing the Target Variant 1062. In some implementations of the technology disclosed, additional variant features can be extracted from Variant Call Format (.vcf) files. The Quality Classifier 1064 is a binary classification model with output classes for High Quality 1066 and Low Quality 1068.

[0093] Figure 11 is an illustrative example 1100 of the variant features in the plurality of variant features describing guanine-cytosine content. A short Genetic Sequence B 1102 contains a proportion of adenine, thymine, guanine, and cytosine nucleic acids. The guanine-cytosine content (GC) of a genetic sequence corresponds to the proportion of guanine and cytosine nucleic acids within the sequence. GC content is a physiochemical descriptor of nucleic acid sequences that can be used as a proxy for thermostability of nucleic acid sequences due to differences in chemical bonding behavior as compared to adenine-thymine bonding behavior. GC content influences read coverage in next-generation sequencing applications. Equation 1122 is used for a sample calculation for the GC content of Genetic Sequence B 1102 wherein GC content is equivalent to the ratio of guanine and cytosine count to the overall count of all nucleic acids. Equation 1124 is used for a sample calculation of genetic skew of Genetic Sequence B 1102 wherein GC skew is determined as the ratio of the difference between guanine count and cytosine count to the sum of guanine count and cytosine count for a given window size. The Window Size Example 1164 illustrates a window size of five. When the window size of five is applied to Genetic Sequence B 1102, GC skew is calculated as shown in Table 1184.

[0094] Figure 12 is an illustrative example 1200 of the variant feature in the plurality of variant features describing local composition complexity. Local composition complexity is a

measure of entropy within a genetic sequence. Genetic Sequence X 1202 contains no variability of nucleic acid composition and therefore has low entropy. Genetic Sequence Z 1242 has high variability of nucleic acid composition and therefore has high entropy. Genetic Sequence Y 1222 contains more variability than Genetic Sequence X 1202 but less than Genetic Sequence Z 1242 therefore it can be described as having a medium (i.e., moderate) level of entropy. Equation 1224 computes the entropy of a genetic sequence in the format of local composition complexity. A sample calculation for Genetic Sequence B 1204 results in an entropy value of 1.92 wherein entropy is equivalent to a sum of logarithmic probabilities scaled by the same respective probability for each nucleic acid.

[0095] Figure 13 is an illustrative example 1300 of the variant feature in the plurality of variant features describing allelic count. Variant One 1302 is shown in shaded grey and Variant Two 1304 is shown in white. Population 1322 contains samples of numerous genetic sequences belonging to either Variant One 1302 or Variant Two 1304. Within Population 1322, there are six total samples belonging to Variant One 1302, thus the total allelic count of Variant One 1302 is six. Within Population 1322, there are nine total samples belonging to Variant Two 1304, thus the total allelic count of Variant One 1304 is nine. The error rate for detecting heterozygote called variants is higher than the comparable error rate of homozygous called variants (i.e., heterozygotic false positives occur at a higher rate than homozygotic false positives).

[0096] Figure 14 is an illustrative example 1400 of the variant feature in the plurality of variant features describing process of mapping sequenced reads to a reference genome wherein an additional step is added to compute a quality metric of the mapping. Sequenced Reads 1402 are mapped to a Reference Genome 1404 to produce a Mapping 1444. Mapping quality scores quantify the likelihood of a misplaced sequenced read to the reference genome. Mapping quality is determined by total possible alignments for a given sequenced read and the count of mismatched base pairs within the alignment. The mapping quality score is reported on a Phred scale, a commonly used logarithmic data scaling technique for error rates in sequencing analysis.

[0097] Figure 15 is an illustrative example 1500 of the variant feature in the plurality of variant features describing strand bias in sequenced reads when mapped to a reference genome. Sequenced Reads 1502 contain reads with different strand orientation (i.e., strands oriented in the 5'→3' direction and strands oriented in the 3'→5' direction). As Sequenced Reads 1502 are mapped to Reference Genome 1504, the Mapping 1544 generated will display sequencing biases based on strand orientation wherein one DNA strand is favored over the other. Strand bias may result in a higher error rate for allele count.

[0098] Figure 16 is an illustrative example 1600 of the variant features in the plurality of variant features describing the depth and coverage of sequenced reads mapped to a reference

genome. Depth and coverage of a particular mapping are measures of mapping quality, where both sequencing coverage and depth of sequencing coverage are proportional metrics to the quality of the particular mapping. Sequenced reads 1602 are mapped to a reference genome 1604 at the various genomic regions along the X-axis. The total percentage of target bases within the reference genome to which sequenced reads are mapped is quantified as the coverage of the genome. The average depth of sequencing coverage is the ratio of the number of reads scaled by read length to the total referenced genome length. This concept is illustrated by visualizing the X-axis as length of the reference genome 1604 with coverage corresponding to the total spread breadth of the aligned sequenced reads 1602, whereas the Y-axis shows visualizes the depth by which the reference genome 1604 is covered.

[0099] Figure 17 is an illustrated flow diagram 1700 representing the variant quality classifier configured as a random forest model to classify a Target Variant 1762 as either belonging to the High Quality Class 1766 or the Low Quality Class 1768. As a quality classifier, the Random Forest Model 1744 takes an Input Target Variant 1762 represented as a vector containing the set of variant features in the plurality of variant features $\{x_1:x_n\}$ where each value of x is a variant feature within the set of variant features in the plurality of variant features describing the Target Variant 1762 and generates a classification from Random Forest Model 1744. In the Random Forest Model 1744, a plurality of decision trees each generate a respective output result of the target variant class and a final result is generated via majority averaging.

[0100] Figure 18 is an illustrated flow diagram 1800 representing the variant quality classifier configured as a logistic regression model to classify a Target Variant 1862 as either belonging to the High Quality Class 1866 or the Low Quality Class 1868. The Quality Classifier 1844 takes an Input Target Variant 1862 represented as a vector containing the set of variant features in the plurality of variant features $\{x_1:x_n\}$ where each value of x is a variant feature within the set of variant features in the plurality of variant features describing the Target Variant 1862 and generates a classification from Logistic Regression Model 1844. In the Logistic Regression Model 1844, the model generates an output value in the range of {0,1} and a decision threshold boundary determines if an input value (i.e., Target Variant 1862) will be classified as an output of 0 or 1 (e.g., a decision threshold boundary of 0.5 leads to values in the range {0,0.4} generating an output of 0 and values in the range {0.5,1} generating an output of 1). Determination of the optimal decision threshold boundary may be determined on the basis of optimization of a particular performance metric when training the logistic regression model, such as accuracy, precision, recall, or a particular error function. The binary output values 0 and 1 are assigned to two output classes, High Quality Class 1866 or Low Quality Class 1868.

[0101] Figure 19 is an illustrated flow diagram 1900 representing the variant quality classifier configured as a neural network to classify a Target Variant 1962 as either belonging to the High Quality Class 1966 or the Low Quality Class 1968. The Quality Classifier 1944 takes an Input Target Variant 1962 represented as a vector containing the set of variant features in the plurality of variant features $\{x_1:x_n\}$ where each value of x is a variant feature within the set of variant features in the plurality of variant features describing the Target Variant 1962 and generates a classification from Neural Network 1944. The neural network model processes the Input Target Variant 1962 via a series of connected layers of nodes which each perform a respective weighted data transformation. Backpropagation through the network updates the weights of each node iteratively during the training process and the final trained model generates an output belonging to the High Quality Class 1966 or Low Quality Class 1968 for the Input Target Variant 1962. At this stage, variants identified as high quality or low quality undergo may further filtering steps in certain embodiments, as described further below.

Unique Mapper

[0102] Figure 20 is a flow diagram 2000 of the unique mapper overview process to further improve the quality of a set of called variants following variant calling or machine learning variant classification. Sequenced Reads of a Sample of a Target Species 2002 undergo a filtering process via Cascading Filters 2004 to remove Low Quality Sequenced Reads 2024. The sequencing data may be obtained from Binary Alignment Map (.bam) files. The series of Cascading Filters 2004 comprise filters that remove variants with incorrect codon match between primates and humans, remove variants with annotation errors, gene-specific filters (*e.g.*, skewed distribution of variant machine learning classifier quality scores compared with exomewide scores or deviations from the Hardy-Weinberg equilibrium), and removal of variants that do not meet a particular machine learning classifier performance metric threshold. The resulting Intermediate Set of Sequenced Reads 2006 is mapped to a Pseudo-Target Reference Genome 2008 and to a Non-Target Reference Genome 2026. The Pseudo-Target Reference Genome 2008 is divided into a number of bins (i.e., sequential nonoverlapping genomic regions of specified equal length). The Non-Target Reference Genome 2026 is also divided into an equivalent number of bins of equivalent size in comparison to the Pseudo-Target Reference Genome 2008 bins. Bins are compared on a one-to-one basis to determine the degree of mapping homology between corresponding bins. The best-mapped bin is identified as the bin wherein the degree of match (i.e., alignment between mapped genome for the bin) and used to generate a Unique Mapper Score 2040. In one embodiment of the technology disclosed, The Unique Mapper Score 2040 is unique for each specific sample, and the Unique Mapper Scores across all samples to a specific

reference target species are averaged to obtain a single mean Unique Mapper Score which applies to all variants of the reference target species that fall into each respective bin.

[0103] Figure 21 schematically illustrates the Gene Annotation Filter 2100 in the series of cascaded filters for variant quality filtering. Gene annotation includes the labeling of a genome for features such as gene location, coding and non-coding regions, and various descriptors of genetic function. Incorrect gene annotation can lead to error in the variant calling process. Gene A 2102 is located at a genomic region and includes Feature X at a specific location within its structure. Gene A 2104 is the correct gene annotation of Gene A 2102 wherein the genomic structure is correctly annotated with Feature X properly located. Gene B 2106 is in a different location and contains different structure (i.e., contains Feature Y rather than Feature X). In a case of incorrect gene annotation by gene prediction error, Gene A 2102 may be incorrectly annotated as Gene B 2106. As a result, any resulting mapping to Gene B 2106 (i.e., despite being labeled as Gene B, the genomic sequence belongs to Gene A) is erroneous. Called variants mapped to genes with annotation errors are filtered out.

[0104] Figure 22 schematically illustrates the process 2200 codon transcription and translation and filtering for codon match. Genetic Sequence A 2202 consists of nucleic acids. The nucleic acid sequence in Genetic Sequence A 2202 undergoes transcription to generate mRNA Transcript A 2242. Following transcription, mRNA Transcript A 2242 is translated to generate Amino Acid Sequence A 2262. Each amino acid is translated from a three nucleic acid sequence referred to as a codon as highlighted by grey shaded boxes for five total codons across Genetic Sequence A 2202, mRNA Transcript A 2242, and Amino Acid Sequence A 2262. Codon B 2282 and Codon C 2284 contain identical nucleic acids and will therefore be transcribed and translated into the same amino acid. If a non-target reference genome contained Codon B 2282 and a pseudo-target reference genome contained Codon C 2284 at the same aligned position, these codons would match and the codon mismatch filter would not remove called variants that align to the genomic region corresponding to Codon B 2282 and Codon C 2284. Codon D 2286 and Codon E 2288 differ in the third nucleic acid position and will not transcribe and translate to the same amino acid. If a non-target reference genome contained Codon D 2286 and a pseudo-target reference genome contained Codon E 2288 at the same aligned position, these aligned codons would not match and called variants that align to the genomic region corresponding to Codon D 2286 and Codon E 2288 are filtered out.

[0105] Figure 23 illustrates the process 2300 of filtering genes based on a distribution of machine learning scores. Scores from a variant quality classifier are plotted on a graph measuring frequency for both a Specific Gene represented by a Specific Gene Distribution 2304 and Exomewide represented by an Exomewide Distribution 2302. A Wilcoxon rank sum test

determines if the Specific Gene Distribution 2304 is skewed in comparison to the Exomewide Distribution 2302 via significance testing for the probability of a randomly selected machine learning score from the Specific Gene Distribution 2304 being greater than a randomly selected machine learning score from the Exomewide Distribution 2302 is equivalent to the probability of a randomly selected machine learning score from the Exomewide Distribution 2302 being greater than a randomly selected machine learning score from the Specific Gene Distribution 2304. Called variants mapped to genes that are determined to be skewed in contrast to the Exomewide Distribution 2302 are filtered out. Determination of skewedness when comparing the Exomewide Distribution 2302 and the Specific Gene Distribution 2304 identifies the gene as an outlier with potential error.

[0106] Figure 24 illustrates deviation from Hardy-Weinberg Equilibrium in an example population 2400. A filter within the Cascading Filters 2004 removes variants that deviate from the Hardy-Weinberg Equilibrium. Dominant alleles are represented by the letter 'p' and recessive alleles are represented by the letter 'q'. Homozygous dominant genotypes (i.e., 'pp'; 2402) are represented by upward diagonal shaded circles. Heterozygous genotypes ('pq'; 2404) are represented by diamond crosshatched circles. Homozygous recessive genotypes ('qq'; 2406) are represented by downward diagonal shaded circles. The population shown includes 25 samples with respective genotypes. In Generation One 2442, each genotype has respective frequencies counted as the proportion of the respective genotype to the total population count. In Generation Two 2444, each genotype has updated respective frequencies counted as the proportion of the respective genotype to the total population count. Populations with unchanged genotype frequencies in sequential generations are considered to be in Hardy-Weinberg Equilibrium. The genotype frequencies for the example population shown in Figure 24 are different in Generation Two 2444 from Generation One 2442 therefore the population deviates from the Hardy-Weinberg Equilibrium. Deviations from the Hardy-Weinberg Equilibrium can result in overcalling of heterozygous genotypes and as a result, called variants mapped to genes that are not in Hardy-Weinberg Equilibrium as determined by large population databases are filtered out.

[0107] Figure 25 is an illustration 2500 of a nonsense variant. A filter within the Cascading Filters 2004 removes nonsense variants. Nonsense variants, also referred to as 'stop gain variants', result from single nucleotide polymorphisms which change a codon sequence such that a previously amino acid-translating codon will translate to a stop codon as a result of the novel mutated amino acid sequence. Premature stop codons prevent the remainder of the mRNA transcript from being translated and as a result the amino acid sequence is terminated early. Genetic Sequence B 2502 is transcribed to mRNA Transcript B 2522, and mRNA Transcript B 2522 is translated into Amino Acid Sequence B 2542 for a total of five codons. Single Nucleotide

Polymorphism 2540 at position 12 results in a change from a guanine nucleic acid to a thymine nucleic acid. As a result, the fourth codon has changed from ACG to ACT and will subsequently be transcribed to a stop codon rather than being transcribed and translated to a cysteine amino acid residue. The premature stop codon ends translation and the fifth codon will never be translated. As further shown by Figure 25, as a result of the Single Nucleotide Polymorphism 2540, Genetic Sequence B 2562 is transcribed to mRNA Transcript B 2582, and mRNA Transcript B 2582 is translated into Amino Acid Sequence B 2510 for a total of four codons, where the fourth codon is now a stop codon.

[0108] Figure 26 contains a graph 2600 of collected results demonstrating the cascading filter effect on number of nonsense variants per sample. In one implementation of the technology disclosed, number of nonsense variants per sample is compared between samples from non-human primate species and human. No filtering of the called variants from samples from non-human primate species results in a significantly higher number of nonsense variants per sample as compared to the corresponding human level of nonsense variants per sample. The called variants from samples from non-human primate species undergo cascaded filters including a codon match filter, gene annotation error filter, machine learning distribution skew filter, Hardy-Weinberg Equilibrium deviation filter, Unique Mapper filter (called variants with Unique Mapper scores less than 0.6 removed), and random forest score filter (called variants with a random forest score greater than 0.17 removed). Boxplots showing the average number of stop-gained variants per sample of each primate reference species was gradually reduced to close to human level after a series of variant filtering steps, including requiring codon-match, removing SNPs in poorly-annotated genes or in genes with skewed random forest (RF) score distribution or deviating from Hardy Weinberg equilibrium, and removing SNPs with unique-mapper score <0.6 or RF score >0.17. Each dot represents the average number of stop-gained variants of each primate reference species. The horizontal line shows the average number of stop-gained variants of human samples from Platinum genome project.

[0109] Figure 27 contains a graph 2700 of collected results demonstrating the cascading filter effect on missense: synonymous ratio of called variants per sample. In one implementation of the technology disclosed, missense: synonymous ratio (MSR; the ratio used to estimate the balance of benign and pathogenic variants present within a particular cohort) is compared between samples from non-human primate species and human. The called variants from samples from non-human primate species undergo cascaded filters including a codon match filter, gene annotation error filter, machine learning distribution skew filter, Hardy-Weinberg Equilibrium deviation filter, Unique Mapper filter (called variants with Unique Mapper scores less than 0.6 removed), and random forest score filter (called variants with a random forest score greater than

0.17 removed). Boxplots showing missense: synonymous ratios decreased after variant filtering steps. Each dot represents the MSR of each primate reference species. The black line represents MSR of human samples.

[0110] Figure 28 contains a graph 2800 of collected results demonstrating the cascading filter effect on number of insertion-deletion variants (indels) per sample. The called variants from samples from non-human primates undergo cascading filters including a gene annotation error filter, machine learning distribution skew filter, Hardy-Weinberg Equilibrium deviation filter, and Unique Mapper filter (called variants with Unique Mapper scores less than 0.6 removed). The average number of indels per sample of each primate reference species diminished after filtering steps.

Computer System

[0111] Figure 29 shows an example computer system 2900 that can be used to implement the technology disclosed. Computer system 2900 includes at least one central processing unit (CPU) 2924 that communicates with a number of peripheral devices via bus subsystem 2922. These peripheral devices can include a storage subsystem 2910 including, for example, memory devices and a file storage subsystem 2918, user interface input devices 2920, user interface output devices 2928, and a network interface subsystem 2926. The input and output devices allow user interaction with computer system 2900. Network interface subsystem 2926 provides an interface to outside networks, including an interface to corresponding interface devices in other computer systems.

[0112] In one implementation, the random forest model 1744 is communicably linked to the storage subsystem 2910 and the user interface input devices 2920.

[0113] User interface input devices 2920 can include a keyboard; pointing devices such as a mouse, trackball, touchpad, or graphics tablet; a scanner; a touch screen incorporated into the display; audio input devices such as voice recognition systems and microphones; and other types of input devices. In general, use of the term “input device” is intended to include all possible types of devices and ways to input information into computer system 2900.

[0114] User interface output devices 2928 can include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices. The display subsystem can include an LED display, a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), a projection device, or some other mechanism for creating a visible image. The display subsystem can also provide a non-visual display such as audio output devices. In general, use of the term “output device” is intended to include all possible types of devices and ways to output information from computer system 2900 to the user or to another machine or computer system.

[0115] Storage subsystem 2910 stores programming and data constructs that provide the functionality of some or all of the modules and methods described herein. These software modules are generally executed by processors 2930.

[0116] Processors 2930 can be graphics processing units (GPUs), field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), and/or coarse-grained reconfigurable architectures (CGRAs). Processors 2930 can be hosted by a deep learning cloud platform such as Google Cloud Platform™, Xilinx™, and Cirrascale™. Examples of processors 2930 include Google's Tensor Processing Unit (TPU)™, rackmount solutions like GX4 Rackmount Series™, GX29 Rackmount Series™, NVIDIA DGX-1™, Microsoft' Stratix V FPGA™, Graphcore's Intelligent Processor Unit (IPU)™, Qualcomm's Zeroth Platform™ with Snapdragon processors™, NVIDIA's Volta™, NVIDIA's DRIVE PX™, NVIDIA's JETSON TX1/TX2 MODULE™, Intel's Nirvana™, Movidius VPU™, Fujitsu DPI™, ARM's DynamicIQ™, IBM TrueNorth™, Lambda GPU Server with Testa V100s™, and others.

[0117] Memory subsystem 2912 used in the storage subsystem 2910 can include a number of memories including a main random access memory (RAM) 2914 for storage of instructions and data during program execution and a read only memory (ROM) 2916 in which fixed instructions are stored. A file storage subsystem 2918 can provide persistent storage for program and data files, and can include a hard disk drive, a floppy disk drive along with associated removable media, a CD-ROM drive, an optical drive, or removable media cartridges. The modules implementing the functionality of certain implementations can be stored by file storage subsystem 2918 in the storage subsystem 2910, or in other machines accessible by the processor.

[0118] Bus subsystem 2922 provides a mechanism for letting the various components and subsystems of computer system 2900 communicate with each other as intended. Although bus subsystem 2922 is shown schematically as a single bus, alternative implementations of the bus subsystem can use multiple busses.

[0119] Computer system 2900 itself can be of varying types including a personal computer, a portable computer, a workstation, a computer terminal, a network computer, a television, a mainframe, a server farm, a widely-distributed set of loosely networked computers, or any other data processing system or user device. Due to the ever-changing nature of computers and networks, the description of computer system 2900 depicted in Figure 29 is intended only as a specific example for purposes of illustrating the preferred implementations of the present invention. Many other configurations of computer system 2900 are possible having more or less components than the computer system depicted in Figure 29.

Clauses

[0120] The technology disclosed, in particularly, the clauses disclosed in this section, can be practiced as a system, method, or article of manufacture. One or more features of an implementation can be combined with the base implementation. Implementations that are not mutually exclusive are taught to be combinable. One or more features of an implementation can be combined with other implementations. This disclosure periodically reminds the user of these options. Omission from some implementations of recitations that repeat these options should not be taken as limiting the combinations taught in the preceding sections – these recitations are hereby incorporated forward by reference into each of the following implementations.

[0121] One or more implementations and clauses of the technology disclosed or elements thereof can be implemented in the form of a computer product, including a non-transitory computer readable storage medium with computer usable program code for performing the method steps indicated. Furthermore, one or more implementations and clauses of the technology disclosed or elements thereof can be implemented in the form of an apparatus including a memory and at least one processor that is coupled to the memory and operative to perform exemplary method steps. Yet further, in another aspect, one or more implementations and clauses of the technology disclosed or elements thereof can be implemented in the form of means for carrying out one or more of the method steps described herein; the means can include (i) hardware module(s), (ii) software module(s) executing on one or more hardware processors, or (iii) a combination of hardware and software modules; any of (i)-(iii) implement the specific techniques set forth herein, and the software modules are stored in a computer readable storage medium (or multiple such media).

[0122] The clauses described in this section can be combined as features. In the interest of conciseness, the combinations of features are not individually enumerated and are not repeated with each base set of features. The reader will understand how features identified in the clauses described in this section can readily be combined with sets of base features identified as implementations in other sections of this application. These clauses are not meant to be mutually exclusive, exhaustive, or restrictive; and the technology disclosed is not limited to these clauses but rather encompasses all possible combinations, modifications, and variations within the scope of the claimed technology and its equivalents.

[0123] Other implementations of the clauses described in this section can include a non-transitory computer readable storage medium storing instructions executable by a processor to perform any of the clauses described in this section. Yet another implementation of the clauses described in this section can include a system including memory and one or more processors

operable to execute instructions, stored in the memory, to perform any of the clauses described in this section.

[0124] We disclose the following clauses:

1. A computer-implemented method of determining feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, including:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

2. The computer-implemented method of clause 1, wherein the pseudo-target species is the target species.

3. The computer-implemented method of clause 1, wherein the pseudo-target species is different from the target species.

4. The computer-implemented method of clause 3, wherein the pseudo-target species is homologous with the target species.

5. The computer-implemented method of clause 1, wherein the non-target species is a human.

6. The computer-implemented method of clause 1, wherein the target species is a non-human primate.

7. The computer-implemented method of clause 1, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

8. The computer-implemented method of clause 1, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

9. The computer-implemented method of clause 1, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared

between the reference genome of the non-target species and the reference genome of the pseudo-target species.

10. The computer-implemented method of clause 1, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

11. The computer-implemented method of clause 10, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

12. A system including one or more processors coupled to memory, the memory loaded with computer instructions to determine feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, the instructions, when executed on the processors, implement actions comprising:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

13. The system of clause 12, wherein the pseudo-target species is the target species.

14. The system of clause 12, wherein the pseudo-target species is different from the target species.

15. The system of clause 12, wherein the pseudo-target species is homologous with the target species.

16. The system of clause 12, wherein the non-target species is a human.

17. The system of clause 12, wherein the target species is a non-human primate.

18. The system of clause 12, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target

species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

19. The system of clause 12, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

20. The system of clause 12, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared between the reference genome of the non-target species and the reference genome of the pseudo-target species.

21. The system of clause 12, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

22. The system of clause 12, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

23. A non-transitory computer readable storage medium impressed with computer program instructions to determine feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, the instructions, when executed on a processor, implement a method comprising:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

24. The non-transitory computer readable storage medium of clause 23, wherein the pseudo-target species is the target species.

25. The non-transitory computer readable storage medium of clause 23, wherein the pseudo-target species is different from the target species.

26. The non-transitory computer readable storage medium of clause 23, wherein the pseudo-target species is homologous with the target species.

27. The non-transitory computer readable storage medium of clause 23, wherein the non-target species is a human.

28. The non-transitory computer readable storage medium of clause 23, wherein the target species is a non-human primate.

29. The non-transitory computer readable storage medium of clause 23, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

30. The non-transitory computer readable storage medium of clause 23, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

31. The non-transitory computer readable storage medium of clause 23, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared between the reference genome of the non-target species and the reference genome of the pseudo-target species.

32. The non-transitory computer readable storage medium of clause 23, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

33. The non-transitory computer readable storage medium of clause 32, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

34. A system, comprising:

a variant quality classifier configured to process a plurality of features of a target variant, and generate a quality indication for the target variant,

wherein the variant quality classifier is trained on a set of high-quality variants and a set of low-quality variants,

wherein high-quality variants in the set of high-quality variants are identified as true positive variants that are common between a first set of variants and a second set of variants,

wherein low-quality variants in the set of low-quality variants are identified as false positive variants that are present in the second set of variants but absent from the first set of variants,

wherein the first set of variants is detected by variant calling sequenced reads of a sample of

a target species against a reference genome of a non-target species, and

wherein the second set of variants is detected by variant calling the sequenced reads of the sample of the target species against a reference genome of a pseudo-target species.

35. The system of clause 34, wherein the variant quality classifier is a random forest model.

36. The system of clause 34, wherein the variant quality classifier is a logistic regression model.

37. The system of clause 34, wherein the variant quality classifier is a neural network model.

38. The system of clause 34, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) content within the sequenced reads of the target variant.

39. The system of clause 34, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) skew within the sequenced reads of the target variant,
wherein the GC skew represents a normalized excess of cytosine over guanine in a given sequenced read of the target variant.

40. The system of clause 34, wherein a feature in the plurality of features of the target variant is a local composition complexity within one hundred base pairs upstream or downstream of the target variant.

41. The system of clause 34, wherein a feature in the plurality of features of the target variant is an allelic count of the sequenced reads of the target variant.

42. The system of clause 34, wherein a feature in the plurality of features of the target variant is a mapping quality of the sequenced reads of the target variant.

43. The system of clause 34, wherein a feature in the plurality of features of the target variant is a p-value of Fisher's exact test to detect strand bias in the sequenced reads of the target variant.

44. The system of clause 34, wherein a feature in the plurality of features of the target variant is a symmetric odds ratio to detect strand bias in the sequenced reads of the target variant.

45. The system of clause 34, wherein a feature in the plurality of features of the target variant is a variant quality by depth of the sequenced reads of the target variant.

46. The system of clause 34, wherein a feature in the plurality of features of the target variant is a genotype quality of the sequenced reads of the target variant.

47. The system of clause 34, wherein a feature in the plurality of features of the target variant is a read depth of the target variant normalized by a mean coverage of the sequenced reads of the target variant.

48. The system of clause 34, wherein a feature in the plurality of features of the target variant is a fraction alternative allele read depth out of a target variant coverage of the sequenced reads of the target variant.

49. The system of clause 34, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within five base pairs upstream or downstream of the sequenced reads of the target variant.

50. The system of clause 34, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within ten base pairs upstream or downstream of the sequenced reads of the target variant.

51. The system of clause 34, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

52. The system of clause 34, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

53. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

54. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

55. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

56. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

57. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

58. The system of clause 34, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

59. A computer-implemented method of processing a plurality of features of a target variant, and generate a quality indication for the target variant, including:

training a variant quality classifier on a set of high-quality variants and a set of low-quality variants;

identifying high-quality variants in the set of high-quality variants as true positive variants that are common between a first set of variants and a second set of variants;

identifying low-quality variants in the set of low-quality variants as false positive variants that are present in the second set of variants but absent from the first set of variants;

detecting the first set of variants by variant calling sequenced reads of a sample of a target species against a reference genome of a non-target species, and

detecting the second set of variants by variant calling the sequenced reads of the sample of the target species against a reference genome of a pseudo-target species.

60. The computer-implemented method of clause 59, wherein the variant quality classifier is a random forest model.

61. The computer-implemented method of clause 59, wherein the variant quality classifier is a logistic regression model.

62. The computer-implemented method of clause 59, wherein the variant quality classifier is a neural network model.

63. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) content within the sequenced reads of the target variant.

64. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) skew within the sequenced reads of the target variant,

wherein the GC skew represents a normalized excess of cytosine over guanine in a given sequenced read of the target variant.

65. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a local composition complexity within one hundred base pairs upstream or downstream of the target variant.

66. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is an allelic count of the sequenced reads of the target variant.

67. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a mapping quality of the sequenced reads of the target variant.

68. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a p-value of Fisher's exact test to detect strand bias in the sequenced reads of the target variant.

69. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a symmetric odds ratio to detect strand bias in the sequenced reads of the target variant.

70. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a variant quality by depth of the sequenced reads of the target variant.

71. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a genotype quality of the sequenced reads of the target variant.

72. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a read depth of the target variant normalized by a mean coverage of the sequenced reads of the target variant.

73. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a fraction alternative allele read depth out of a target variant coverage of the sequenced reads of the target variant.

74. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within five base pairs upstream or downstream of the sequenced reads of the target variant.

75. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within ten base pairs upstream or downstream of the sequenced reads of the target variant.

76. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

77. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

78. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

79. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

80. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

81. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

82. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

83. The computer-implemented method of clause 59, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

84. A non-transitory computer readable storage medium impressed with computer program instructions to process a plurality of features of a target variant, and generate a quality indication for the target variant, the instructions, when executed on a processor, implement a method comprising:

a variant quality classifier trained on a set of high-quality variants and a set of low-quality

variants,

wherein high-quality variants in the set of high-quality variants are identified as true positive variants that are common between a first set of variants and a second set of variants,

wherein low-quality variants in the set of low-quality variants are identified as false positive variants that are present in the second set of variants but absent from the first set of variants,

wherein the first set of variants is detected by variant calling sequenced reads of a sample of a target species against a reference genome of a non-target species, and

wherein the second set of variants is detected by variant calling the sequenced reads of the sample of the target species against a reference genome of a pseudo-target species.

85. The non-transitory computer readable storage medium of clause 84, wherein the variant quality classifier is a random forest model.

86. The non-transitory computer readable storage medium of clause 84, wherein the variant quality classifier is a logistic regression model.

87. The non-transitory computer readable storage medium of clause 84, wherein the variant quality classifier is a neural network model.

88. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) content within the sequenced reads of the target variant.

89. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) skew within the sequenced reads of the target variant,

wherein the GC skew represents a normalized excess of cytosine over guanine in a given sequenced read of the target variant.

90. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a local composition complexity within one hundred base pairs upstream or downstream of the target variant.

91. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is an allelic count of the sequenced reads of the target variant.

92. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a mapping quality of the sequenced reads of the target variant.

93. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a p-value of Fisher's exact test to detect strand bias in the sequenced reads of the target variant.

94. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a symmetric odds ratio to detect strand bias in the sequenced reads of the target variant.

95. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a variant quality by depth of the sequenced reads of the target variant.

96. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a genotype quality of the sequenced reads of the target variant.

97. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a read depth of the target variant normalized by a mean coverage of the sequenced reads of the target variant.

98. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a fraction alternative allele read depth out of a target variant coverage of the sequenced reads of the target variant.

99. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within five base pairs upstream or downstream of the sequenced reads of the target variant.

100. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within ten base pairs upstream or downstream of the sequenced reads of the target variant.

101. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

102. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

103. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of

the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

104. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

105. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

106. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

107. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

108. The non-transitory computer readable storage medium of clause 84, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

109. A computer-implemented method of identifying and excluding regions that do not have one-to-one mapping between a first reference genome and a second reference genome, including:

accessing sequenced reads of a sample of a target species;

identifying and removing, from the sequenced reads, low-quality sequenced reads based on applying a mapping quality filter to the sequenced reads, thereby pruning high-quality sequenced reads from the sequenced reads;

segmenting a non-target reference genome of a non-target species into a plurality of bins, and then, on a bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins

in the non-target reference genome;

segmenting a pseudo-target reference genome of a pseudo-target species into the plurality of bins, and then, on the bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the pseudo-target reference genome;

identifying a best-mapped bin in the pseudo-target reference genome based on a greatest degree of match between the best-mapped bin in the pseudo-target reference genome and a corresponding bin in the non-target reference genome, wherein a degree of match between corresponding bins in the pseudo-target reference genome and the non-target reference genome is determined by a number of reads mapped between the corresponding bins;

generating a unique-mapper score for the pseudo-target reference genome based on a number of reads mapped between the best-mapped bin in the pseudo-target reference genome and the corresponding bin in the non-target reference genome; and

using the unique-mapper score to identify and exclude low-quality sequenced reads.

110. The computer-implemented method of clause 109, wherein low-quality sequenced reads include stop-gained variants.

111. The computer-implemented method of clause 109, wherein a plurality of cascading filters the applied to the sequenced reads of the sample of the target species to filter low-quality sequenced reads.

112. The computer-implemented method of clause 111, wherein a filter within the plurality of cascading filters is configured to detect and exclude genetic regions possessing incorrect gene annotation in a reference genome.

113. The computer-implemented method of clause 111, wherein a filter within the plurality of cascading filters is configured to detect and exclude codons that do not match between the pseudo-target species reference genome and the non-target species reference genome.

114. The computer-implemented method of clause 111, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing a skewed distribution of variant classifier scores in comparison to a distribution of variant classifier scores for the complete reference genome.

115. The computer-implemented method of clause 111, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing deviations from the Hardy-Weinberg equilibrium.

116. The computer-implemented method of clause 111, wherein a filter within the plurality of cascading filters is configured to detect and remove single nucleotide polymorphisms with a random forest score of greater than 0.17.

117. The computer-implemented method of clause 109, wherein a one-to-one mapping describes the fraction of a number of reads in one bin within non-target reference genome map to a single corresponding region within a pseudo-target reference genome.

118. The computer-implemented method of clause 117, wherein consecutive identical bins are collectively considered as a single bin and do not exclude the possibility of a one-to-one mapping.

119. The computer-implemented method of clause 117, wherein more than two nonconsecutive identical bins are considered duplicate regions and exclude the possibility of a one-to-one mapping.

120. The computer-implemented method of clause 109, wherein a bin describes a one kilobase (kb) region within a reference genome.

121. The computer-implemented method of clause 109, wherein a fraction of reads mapped is detected for the best-mapped region as determined by the mapper score.

122. The computer-implemented method of clause 109, wherein the unique mapper score is determined by an average of a top fraction across samples for each reference genome.

123. The computer-implemented method of clause 109, wherein the unique mapper score is configured as a filter to exclude sequenced reads with a mapper score less than 20.

124. The computer-implemented method of clause 109, wherein the pseudo-target species is a human.

125. The computer-implemented method of clause 109, wherein the pseudo-target species is a non-human primate.

126. The computer-implemented method of clause 109, wherein the non-target species is a human.

127. The computer-implemented method of clause 109, wherein the non-target species is a non-human primate.

128. The computer-implemented method of clause 109, wherein the target species is a human.

129. The computer-implemented method of clause 109, wherein the target species is a non-human primate.

130. The computer-implemented method of clause 109, wherein the target species and non-target species are homologous.

131. The computer-implemented method of clause 109, wherein the target species and pseudo-target species are homologous.

132. The computer-implemented method of clause 109, wherein the quality of a variant identified from sequenced reads of a target species is a proxy for evolutionary constraint on the variant gene.

133. A system including one or more processors coupled to memory, the memory loaded with computer instructions to identify and exclude regions that do not have one-to-one mapping between a first reference genome and a second reference genome, the instructions, when executed on the processors, implement actions comprising:

accessing sequenced reads of a sample of a target species;

identifying and removing, from the sequenced reads, low-quality sequenced reads based on applying a mapping quality filter to the sequenced reads, thereby pruning high-quality sequenced reads from the sequenced reads;

segmenting a non-target reference genome of a non-target species into a plurality of bins, and then, on a bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the non-target reference genome;

segmenting a pseudo-target reference genome of a pseudo-target species into the plurality of bins, and then, on the bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the pseudo-target reference genome;

identifying a best-mapped bin in the pseudo-target reference genome based on a greatest degree of match between the best-mapped bin in the pseudo-target reference genome and a corresponding bin in the non-target reference genome, wherein a degree of match between corresponding bins in the pseudo-target reference genome and the non-target reference genome is determined by a number of reads mapped between the corresponding bins;

generating a unique-mapper score for the pseudo-target reference genome based on a number of reads mapped between the best-mapped bin in the pseudo-target reference genome and the corresponding bin in the non-target reference genome; and

using the unique-mapper score to identify and exclude low-quality sequenced reads.

134. The system of clause 133, wherein low-quality sequenced reads include stop-gained variants.

135. The system of clause 133, wherein a plurality of cascading filters the applied to the sequenced reads of the sample of the target species to filter low-quality sequenced reads.

136. The system of clause 135, wherein a filter within the plurality of cascading filters is configured to detect and exclude genetic regions possessing incorrect gene annotation in a reference genome.

137. The system of clause 135, wherein a filter within the plurality of cascading filters is configured to detect and exclude codons that do not match between the pseudo-target species reference genome and the non-target species reference genome.

138. The system of clause 135, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing a skewed

distribution of variant classifier scores in comparison to a distribution of variant classifier scores for the complete reference genome.

139. The system of clause 135, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing deviations from the Hardy-Weinberg equilibrium.

140. The system of clause 135, wherein a filter within the plurality of cascading filters is configured to detect and remove single nucleotide polymorphisms with a random forest score of greater than 0.17.

141. The system of clause 133, wherein a one-to-one mapping describes the fraction of a number of reads in one bin within non-target reference genome map to a single corresponding region within a pseudo-target reference genome.

142. The system of clause 141, wherein consecutive identical bins are collectively considered as a single bin and do not exclude the possibility of a one-to-one mapping.

143. The system of clause 141, wherein more than two nonconsecutive identical bins are considered duplicate regions and exclude the possibility of a one-to-one mapping.

144. The system of clause 133, wherein a bin describes a one kilobase (kb) region within a reference genome.

145. The system of clause 133, wherein a fraction of reads mapped is detected for the best-mapped region as determined by the mapper score.

146. The system of clause 133, wherein the unique mapper score is determined by an average of a top fraction across samples for each reference genome.

147. The system of clause 133, wherein the unique mapper score is configured as a filter to exclude sequenced reads with a mapper score less than 20.

148. The system of clause 133, wherein the pseudo-target species is a human.

149. The system of clause 133, wherein the pseudo-target species is a non-human primate.

150. The system of clause 133, wherein the non-target species is a human.

151. The system of clause 133, wherein the non-target species is a non-human primate.

152. The system of clause 133, wherein the target species is a human.

153. The system of clause 133, wherein the target species is a non-human primate.

154. The system of clause 133, wherein the target species and non-target species are homologous.

155. The system of clause 133, wherein the target species and pseudo-target species are homologous.

156. The system of clause 133, wherein the quality of a variant identified from sequenced reads of a target species is a proxy for evolutionary constraint on the variant gene.

157. A non-transitory computer readable storage medium impressed with computer program instructions to identify and exclude regions that do not have one-to-one mapping between a first reference genome and a second reference genome, the instructions, when executed on a processor, implement a method comprising:

accessing sequenced reads of a sample of a target species;

identifying and removing, from the sequenced reads, low-quality sequenced reads based on applying a mapping quality filter to the sequenced reads, thereby pruning high-quality sequenced reads from the sequenced reads;

segmenting a non-target reference genome of a non-target species into a plurality of bins, and then, on a bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the non-target reference genome;

segmenting a pseudo-target reference genome of a pseudo-target species into the plurality of bins, and then, on the bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the pseudo-target reference genome;

identifying a best-mapped bin in the pseudo-target reference genome based on a greatest degree of match between the best-mapped bin in the pseudo-target reference genome and a corresponding bin in the non-target reference genome, wherein a degree of match between corresponding bins in the pseudo-target reference genome and the non-target reference genome is determined by a number of reads mapped between the corresponding bins;

generating a unique-mapper score for the pseudo-target reference genome based on a number of reads mapped between the best-mapped bin in the pseudo-target reference genome and the corresponding bin in the non-target reference genome; and

using the unique-mapper score to identify and exclude low-quality sequenced reads.

158. The non-transitory computer readable storage medium of clause 157, wherein low-quality sequenced reads include stop-gained variants.

159. The non-transitory computer readable storage medium of clause 157, wherein a plurality of cascading filters the applied to the sequenced reads of the sample of the target species to filter low-quality sequenced reads.

160. The non-transitory computer readable storage medium of clause 159, wherein a filter within the plurality of cascading filters is configured to detect and exclude genetic regions possessing incorrect gene annotation in a reference genome.

161. The non-transitory computer readable storage medium of clause 159, wherein a filter within the plurality of cascading filters is configured to detect and exclude codons that do not match between the pseudo-target species reference genome and the non-target species reference genome.

162. The non-transitory computer readable storage medium of clause 159, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing a skewed distribution of variant classifier scores in comparison to a distribution of variant classifier scores for the complete reference genome.

163. The non-transitory computer readable storage medium of clause 159, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing deviations from the Hardy-Weinberg equilibrium.

164. The non-transitory computer readable storage medium of clause 159, wherein a filter within the plurality of cascading filters is configured to detect and remove single nucleotide polymorphisms with a random forest score of greater than 0.17.

165. The non-transitory computer readable storage medium of clause 157, wherein a one-to-one mapping describes the fraction of a number of reads in one bin within non-target reference genome map to a single corresponding region within a pseudo-target reference genome.

166. The non-transitory computer readable storage medium of clause 165, wherein consecutive identical bins are collectively considered as a single bin and do not exclude the possibility of a one-to-one mapping.

167. The non-transitory computer readable storage medium of clause 165, wherein more than two nonconsecutive identical bins are considered duplicate regions and exclude the possibility of a one-to-one mapping.

168. The non-transitory computer readable storage medium of clause 157, wherein a bin describes a one kilobase (kb) region within a reference genome.

169. The non-transitory computer readable storage medium of clause 157, wherein a fraction of reads mapped is detected for the best-mapped region as determined by the mapper score.

170. The non-transitory computer readable storage medium of clause 157, wherein the unique mapper score is determined by an average of a top fraction across samples for each reference genome.

171. The non-transitory computer readable storage medium of clause 157, wherein the unique mapper score is configured as a filter to exclude sequenced reads with a mapper score less than 20.

172. The non-transitory computer readable storage medium of clause 157, wherein the pseudo-target species is a human.

173. The non-transitory computer readable storage medium of clause 157, wherein the pseudo-target species is a non-human primate.

174. The non-transitory computer readable storage medium of clause 157, wherein the non-target species is a human.

175. The non-transitory computer readable storage medium of clause 157, wherein the non-target species is a non-human primate.

176. The non-transitory computer readable storage medium of clause 157, wherein the target species is a human.

177. The non-transitory computer readable storage medium of clause 157, wherein the target species is a non-human primate.

178. The non-transitory computer readable storage medium of clause 157, wherein the target species and non-target species are homologous.

179. The non-transitory computer readable storage medium of clause 157, wherein the target species and pseudo-target species are homologous.

180. The non-transitory computer readable storage medium of clause 157, wherein the quality of a variant identified from sequenced reads of a target species is a proxy for evolutionary constraint on the variant gene.

[0125] What we claim is:

CLAIMS

1. A computer-implemented method of determining feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, including:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

2. The computer-implemented method of claim 1, wherein the pseudo-target species is the target species.

3. The computer-implemented method of claim 1, wherein the pseudo-target species is different from the target species.

4. The computer-implemented method of claim 3, wherein the pseudo-target species is homologous with the target species.

5. The computer-implemented method of claim 1, wherein the non-target species is a human.

6. The computer-implemented method of claim 1, wherein the target species is a non-human primate.

7. The computer-implemented method of claim 1, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

8. The computer-implemented method of claim 1, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

9. The computer-implemented method of claim 1, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared

between the reference genome of the non-target species and the reference genome of the pseudo-target species.

10. The computer-implemented method of claim 1, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

11. The computer-implemented method of claim 10, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

12. A system including one or more processors coupled to memory, the memory loaded with computer instructions to determine feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, the computer instructions, when executed on the one or more processors, implement actions comprising:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

13. The system of claim 12, wherein the pseudo-target species is the target species.

14. The system of claim 12, wherein the pseudo-target species is different from the target species.

15. The system of claim 12, wherein the pseudo-target species is homologous with the target species.

16. The system of claim 12, wherein the non-target species is a human.

17. The system of claim 12, wherein the target species is a non-human primate.

18. The system of claim 12, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

19. The system of claim 12, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

20. The system of claim 12, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared between the reference genome of the non-target species and the reference genome of the pseudo-target species.

21. The system of claim 12, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

22. The system of claim 21, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

23. A non-transitory computer readable storage medium impressed with computer program instructions to determine feasibility of using a reference genome of a non-target species for variant calling a sample of a target species, the computer instructions, when executed on a processor, implement a method comprising:

mapping sequenced reads of a sample of a target species to a reference genome of a non-target species to detect a first set of variants in the sequenced reads of the sample of the target species;

mapping the sequenced reads of the sample of the target species to a reference genome of a pseudo-target species to detect a second set of variants in the sequenced reads of the sample of the target species;

comparing the first set of variants and the second set of variants, and identifying a subset of true positive variants that are common between the first set of variants and the second set of variants;

comparing the first set of variants and the second set of variants, and identifying a subset of false positive variants that are present in the second set of variants but absent from the first set of variants; and

based on a count of the subset of false positive variants determining the feasibility of using the reference genome of the non-target species for variant calling the target species.

24. The non-transitory computer readable storage medium of claim 23, wherein the pseudo-target species is the target species.

25. The non-transitory computer readable storage medium of claim 23, wherein the pseudo-target species is different from the target species.

26. The non-transitory computer readable storage medium of claim 23, wherein the pseudo-target species is homologous with the target species.

27. The non-transitory computer readable storage medium of claim 23, wherein the non-target species is a human.

28. The non-transitory computer readable storage medium of claim 23, wherein the target species is a non-human primate.

29. The non-transitory computer readable storage medium of claim 23, further including detecting the second set of variants by mapping the sequenced reads of the sample of the target species to the reference genome of the target species, and then lifting-over the mapped sequenced reads of the sample of the target species to the reference genome of the non-target species.

30. The non-transitory computer readable storage medium of claim 23, further including applying a first filter to filter out low-quality variants from the first set of variants and the second set of variants.

31. The non-transitory computer readable storage medium of claim 23, further including applying a second filter to filter out, from the first set of variants and the second set of variants, fixed substitutions shared between the reference genome of the non-target species and the reference genome of the pseudo-target species.

32. The non-transitory computer readable storage medium of claim 23, wherein false positive variants in the subset of false positive variants occur because a particular region in the sequenced reads of the sample of the target species map to a first region in the reference genome of the non-target species and a second region in the reference genome of the pseudo-target species, wherein the first region and the second region are different.

33. The non-transitory computer readable storage medium of claim 32, wherein the false positive variants occur because the particular region in the sequenced reads of the sample of the target species maps multiple regions in the reference genome of the non-target species.

34. A system, comprising:

a variant quality classifier configured to process a plurality of features of a target variant, and generate a quality indication for the target variant,

wherein the variant quality classifier is trained on a set of high-quality variants and a set of low-quality variants,

wherein high-quality variants in the set of high-quality variants are identified as true positive

variants that are common between a first set of variants and a second set of variants,

wherein low-quality variants in the set of low-quality variants are identified as false positive variants that are present in the second set of variants but absent from the first set of variants,

wherein the first set of variants is detected by variant calling sequenced reads of a sample of a target species against a reference genome of a non-target species, and

wherein the second set of variants is detected by variant calling the sequenced reads of the sample of the target species against a reference genome of a pseudo-target species.

35. The system of claim 34, wherein the variant quality classifier is a random forest model.

36. The system of claim 34, wherein the variant quality classifier is a logistic regression model.

37. The system of claim 34, wherein the variant quality classifier is a neural network model.

38. The system of claim 34, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) content within the sequenced reads of the target variant.

39. The system of claim 34, wherein a feature in the plurality of features of the target variant is a guanine-cytosine (GC) skew within the sequenced reads of the target variant,

wherein the GC skew represents a normalized excess of cytosine over guanine in a given sequenced read of the target variant.

40. The system of claim 34, wherein a feature in the plurality of features of the target variant is a local composition complexity within one hundred base pairs upstream or downstream of the target variant.

41. The system of claim 34, wherein a feature in the plurality of features of the target variant is an allelic count of the sequenced reads of the target variant.

42. The system of claim 34, wherein a feature in the plurality of features of the target variant is a mapping quality of the sequenced reads of the target variant.

43. The system of claim 34, wherein a feature in the plurality of features of the target variant is a p-value of Fisher's exact test to detect strand bias in the sequenced reads of the target variant.

44. The system of claim 34, wherein a feature in the plurality of features of the target variant is a symmetric odds ratio to detect strand bias in the sequenced reads of the target variant.

45. The system of claim 34, wherein a feature in the plurality of features of the target variant is a variant quality by depth of the sequenced reads of the target variant.

46. The system of claim 34, wherein a feature in the plurality of features of the target variant is a genotype quality of the sequenced reads of the target variant.

47. The system of claim 34, wherein a feature in the plurality of features of the target variant is a read depth of the target variant normalized by a mean coverage of the sequenced reads of the target variant.

48. The system of claim 34, wherein a feature in the plurality of features of the target variant is a fraction alternative allele read depth out of a target variant coverage of the sequenced reads of the target variant.

49. The system of claim 34, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within five base pairs upstream or downstream of the sequenced reads of the target variant.

50. The system of claim 34, wherein a feature in the plurality of features of the target variant is an existence of insertion and/or deletion (indel) mutations within ten base pairs upstream or downstream of the sequenced reads of the target variant.

51. The system of claim 34, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

52. The system of claim 34, wherein a feature in the plurality of features of the target variant is a mean coverage of flanking regions five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by the mean coverage of the sequenced reads of the target variant.

53. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

54. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of heterozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

55. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within one hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

56. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

57. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within one hundred base pairs

upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

58. The system of claim 34, wherein a feature in the plurality of features of the target variant is a number of alternate homozygote single nucleotide polymorphisms within five hundred base pairs upstream or downstream of the sequenced reads of the target variant normalized by a median count of variants within the same length regions of the sequenced reads of the target variant.

59. A computer-implemented method of identifying and excluding regions that do not have one-to-one mapping between a first reference genome and a second reference genome, including:

accessing sequenced reads of a sample of a target species;

identifying and removing, from the sequenced reads, low-quality sequenced reads based on applying a mapping quality filter to the sequenced reads, thereby pruning high-quality sequenced reads from the sequenced reads;

segmenting a non-target reference genome of a non-target species into a plurality of bins, and then, on a bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the non-target reference genome;

segmenting a pseudo-target reference genome of a pseudo-target species into the plurality of bins, and then, on the bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the pseudo-target reference genome;

identifying a best-mapped bin in the pseudo-target reference genome based on a greatest degree of match between the best-mapped bin in the pseudo-target reference genome and a corresponding bin in the non-target reference genome, wherein a degree of match between corresponding bins in the pseudo-target reference genome and the non-target reference genome is determined by a number of reads mapped between the corresponding bins;

generating a unique-mapper score for the pseudo-target reference genome based on a number of reads mapped between the best-mapped bin in the pseudo-target reference genome and the corresponding bin in the non-target reference genome; and

using the unique-mapper score to identify and exclude low-quality sequenced reads.

60. The computer-implemented method of claim 59, wherein low-quality sequenced reads include stop-gained variants.

61. The computer-implemented method of claim 59, wherein a plurality of cascading filters are applied to the sequenced reads of the sample of the target species to filter low-quality sequenced reads.

62. The computer-implemented method of claim 61, wherein a filter within the plurality of cascading filters is configured to detect and exclude genetic regions possessing incorrect gene annotation in a reference genome.

63. The computer-implemented method of claim 61, wherein a filter within the plurality of cascading filters is configured to detect and exclude codons that do not match between the pseudo-target species reference genome and the non-target species reference genome.

64. The computer-implemented method of claim 61, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing a skewed distribution of variant classifier scores in comparison to a distribution of variant classifier scores for a complete reference genome.

65. The computer-implemented method of claim 61, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing deviations from a Hardy-Weinberg equilibrium.

66. The computer-implemented method of claim 61, wherein a filter within the plurality of cascading filters is configured to detect and remove single nucleotide polymorphisms with a random forest score of greater than 0.17.

67. The computer-implemented method of claim 59, wherein a one-to-one mapping describes a fraction of a number of reads in one bin within non-target reference genome map to a single corresponding region within a pseudo-target reference genome.

68. The computer-implemented method of claim 67, wherein consecutive identical bins are collectively considered as a single bin and do not exclude a possibility of a one-to-one mapping.

69. The computer-implemented method of claim 67, wherein more than two nonconsecutive identical bins are considered duplicate regions and exclude a possibility of a one-to-one mapping.

70. The computer-implemented method of claim 59, wherein a bin describes a one kilobase (kb) region within a reference genome.

71. The computer-implemented method of claim 59, wherein a fraction of reads mapped is detected for a best-mapped region as determined by the unique-mapper score.

72. The computer-implemented method of claim 59, wherein the unique-mapper score is determined by an average of a top fraction across samples for each reference genome.

73. The computer-implemented method of claim 59, wherein the unique-mapper score is configured as a filter to exclude sequenced reads with a mapper score less than 20.

74. The computer-implemented method of claim 59, wherein the pseudo-target species is a human.

75. The computer-implemented method of claim 59, wherein the pseudo-target species is a non-human primate.

76. The computer-implemented method of claim 59, wherein the non-target species is a human.

77. The computer-implemented method of claim 59, wherein the non-target species is a non-human primate.

78. The computer-implemented method of claim 59, wherein the target species is a human.

79. The computer-implemented method of claim 59, wherein the target species is a non-human primate.

80. The computer-implemented method of claim 59, wherein the target species and non-target species are homologous.

81. The computer-implemented method of claim 59, wherein the target species and pseudo-target species are homologous.

82. The computer-implemented method of claim 59, wherein a quality of a variant identified from sequenced reads of a target species is a proxy for evolutionary constraint on a variant gene.

83. A system including one or more processors coupled to memory, the memory loaded with computer instructions to identify and exclude regions that do not have one-to-one mapping between a first reference genome and a second reference genome, the computer instructions, when executed on the one or more processors, implement actions comprising:

accessing sequenced reads of a sample of a target species;

identifying and removing, from the sequenced reads, low-quality sequenced reads based on applying a mapping quality filter to the sequenced reads, thereby pruning high-quality sequenced reads from the sequenced reads;

segmenting a non-target reference genome of a non-target species into a plurality of bins, and then, on a bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the non-target reference genome;

segmenting a pseudo-target reference genome of a pseudo-target species into the plurality of bins, and then, on the bin-by-bin basis, mapping the high-quality sequenced reads to the plurality of bins in the pseudo-target reference genome;

identifying a best-mapped bin in the pseudo-target reference genome based on a greatest degree of match between the best-mapped bin in the pseudo-target reference genome and a corresponding bin in the non-target reference genome, wherein a degree of match between corresponding bins in the pseudo-target reference genome and the non-target reference genome is determined by a number of reads mapped between the corresponding bins;

generating a unique-mapper score for the pseudo-target reference genome based on a number of reads mapped between the best-mapped bin in the pseudo-target reference genome and the corresponding bin in the non-target reference genome; and

using the unique-mapper score to identify and exclude low-quality sequenced reads.

84. The system of claim 83, wherein low-quality sequenced reads include stop-gained variants.

85. The system of claim 83, wherein a plurality of cascading filters are applied to the sequenced reads of the sample of the target species to filter low-quality sequenced reads.

86. The system of claim 85, wherein a filter within the plurality of cascading filters is configured to detect and exclude genetic regions possessing incorrect gene annotation in a reference genome.

87. The system of claim 85, wherein a filter within the plurality of cascading filters is configured to detect and exclude codons that do not match between the pseudo-target species reference genome and the non-target species reference genome.

88. The system of claim 85, wherein a filter within the plurality of cascading filters is configured to detect and exclude genes within a reference genome possessing a skewed distribution of variant classifier scores in comparison to a distribution of variant classifier scores for a complete reference genome.

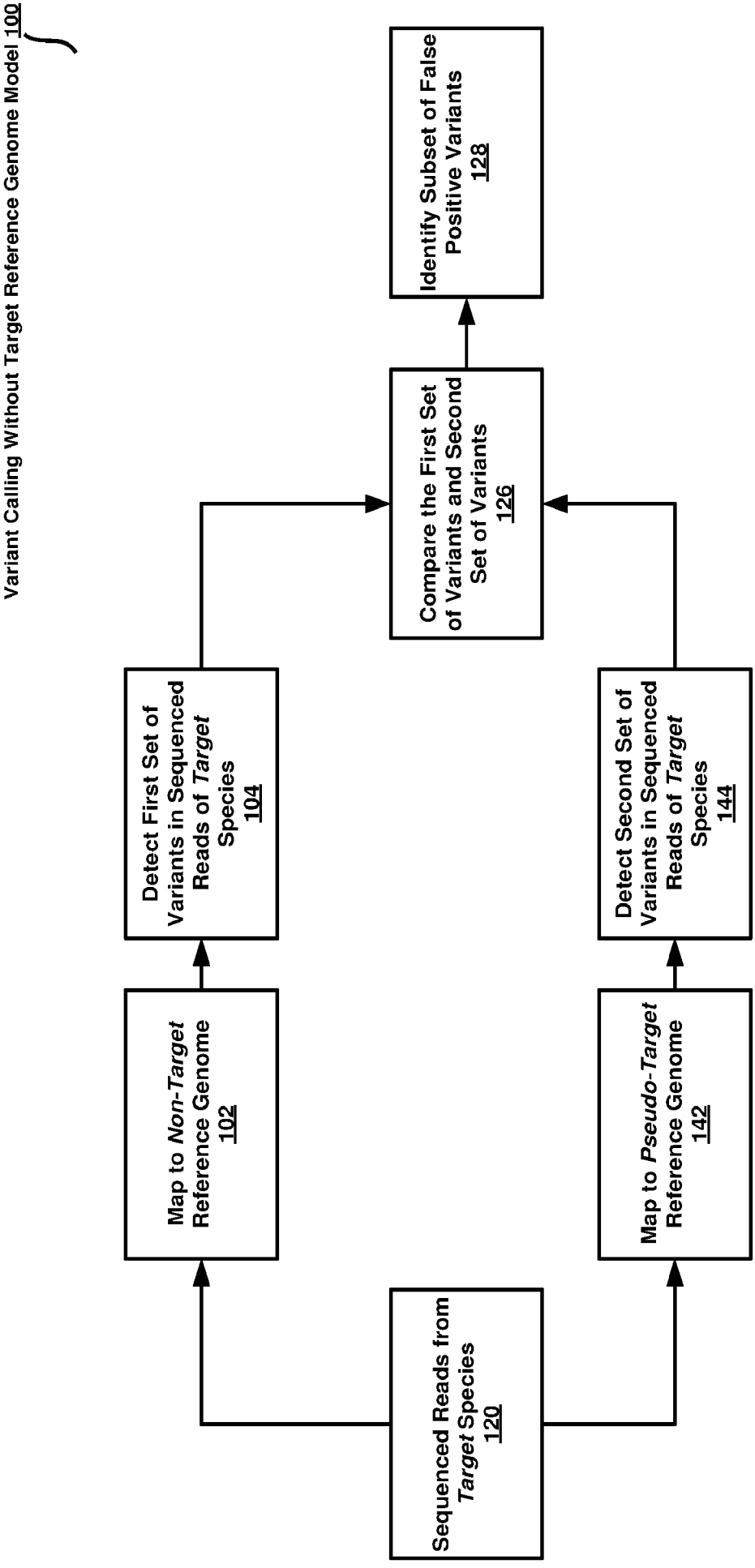

Variant Calling Without Target Reference Genome Model 100
Sequenced Reads from Target Species 120
Map to Non-Target Reference Genome 102
Map to Pseudo-Target Reference Genome 142
Detect First Set of Variants in Sequenced Reads of Target Species 104
Detect Second Set of Variants in Sequenced Reads of Target Species 144
Compare the First Set of Variants and Second Set of Variants 126
Identify Subset of False Positive Variants 128

Figure 1

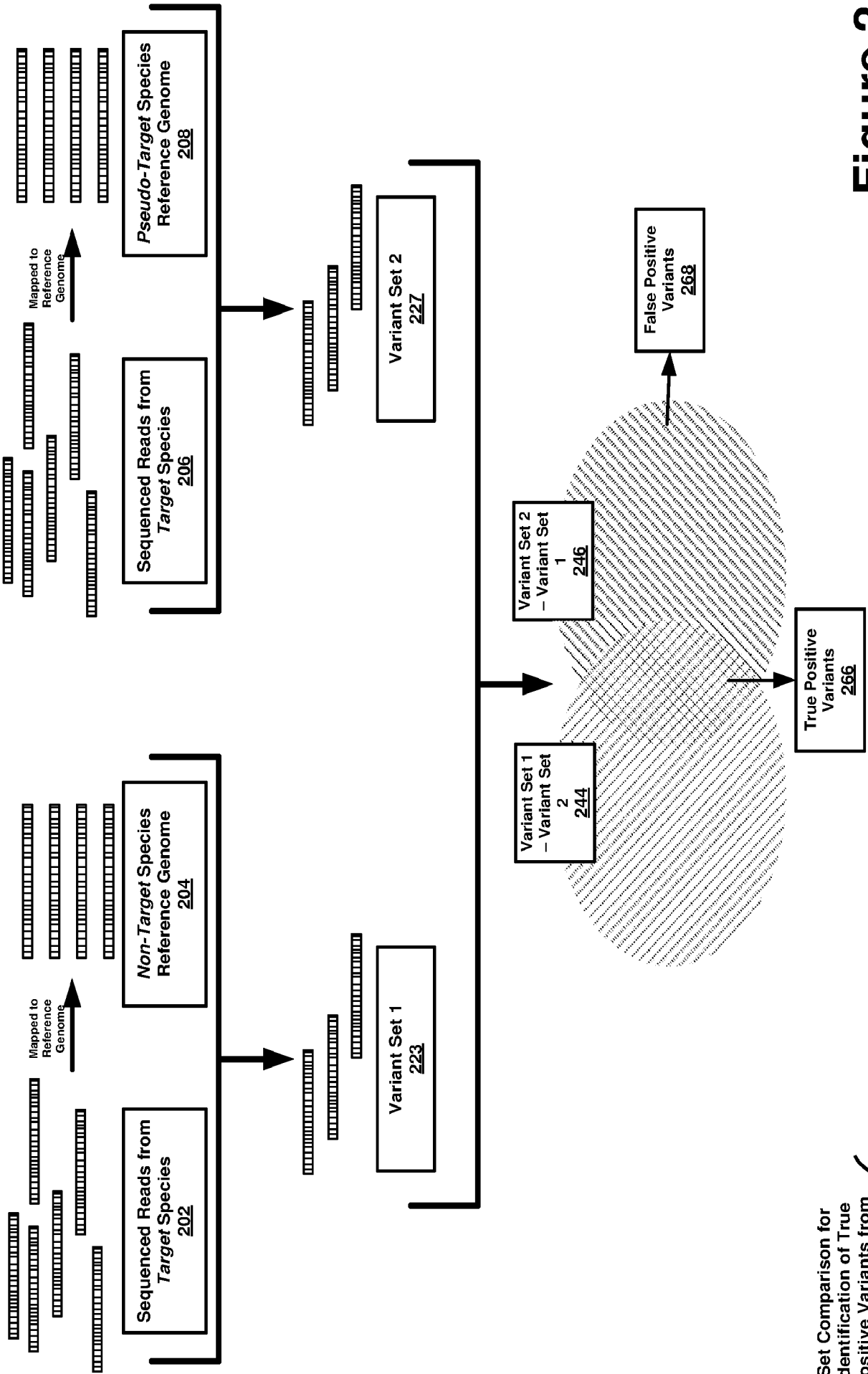

Sequenced Reads from Target Species 202
Mapped to Reference Genome
Non-Target Species Reference Genome 204
Variant Set 1 223
Sequenced Reads from Target Species 206
Mapped to Reference Genome
Pseudo-Target Species Reference Genome 208
Variant Set 2 227
Variant Set 1 – Variant Set 2 244
Variant Set 2 – Variant Set 1 246
True Positive Variants 266
False Positive Variants 268
Set Comparison for Identification of True Positive Variants from Sequenced Reads 200

Figure 2

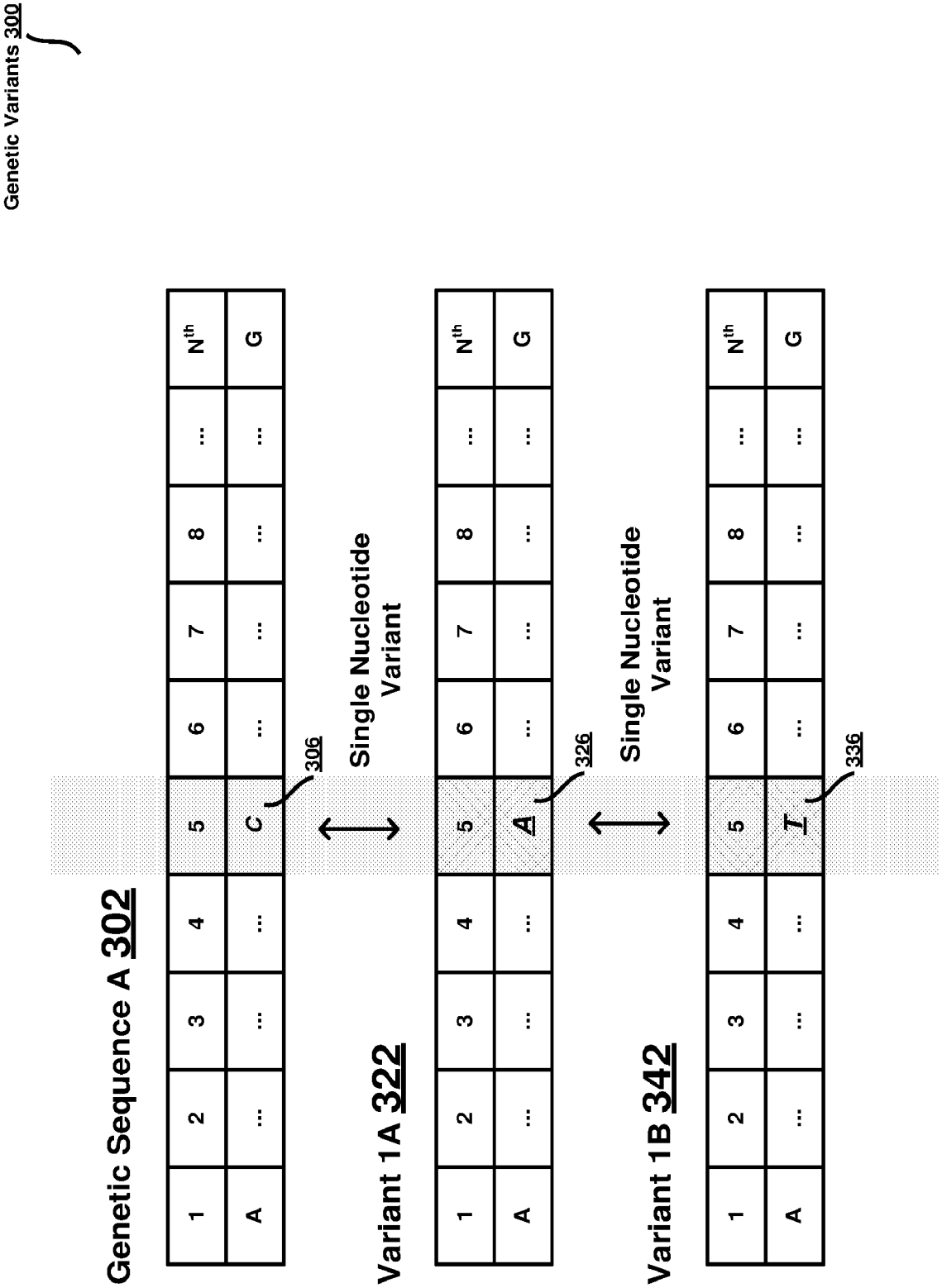

Genetic Variants 300
Genetic Sequence A 302
1 2 3 4 5 6 7 8 ... Nth
A ... ... C ... ... G
306
Single Nucleotide Variant
Variant 1A 322
1 2 3 4 5 6 7 8 ... Nth
A ... ... A ... ... G
326
Single Nucleotide Variant
Variant 1B 342
1 2 3 4 5 6 7 8 ... Nth
A ... ... T ... ... G
336

Figure 3

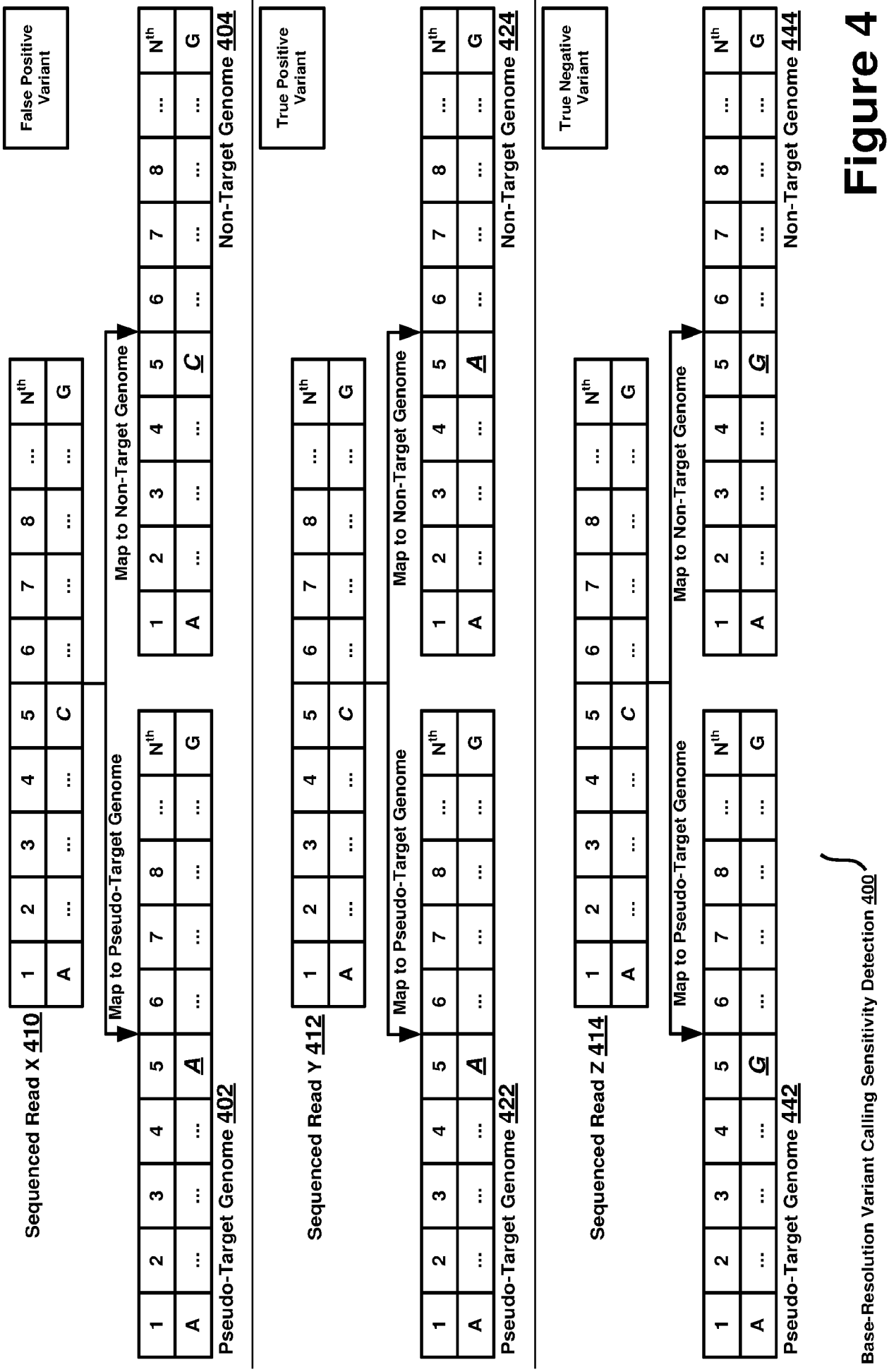

False Positive Variant
Sequenced Read X 410
Map to Pseudo-Target Genome
Pseudo-Target Genome 402
Map to Non-Target Genome
Non-Target Genome 404
True Positive Variant
Sequenced Read Y 412
Map to Pseudo-Target Genome
Pseudo-Target Genome 422
Map to Non-Target Genome
Non-Target Genome 424
True Negative Variant
Sequenced Read Z 414
Map to Pseudo-Target Genome
Pseudo-Target Genome 442
Map to Non-Target Genome
Non-Target Genome 444
Base-Resolution Variant Calling Sensitivity Detection 400

Figure 4

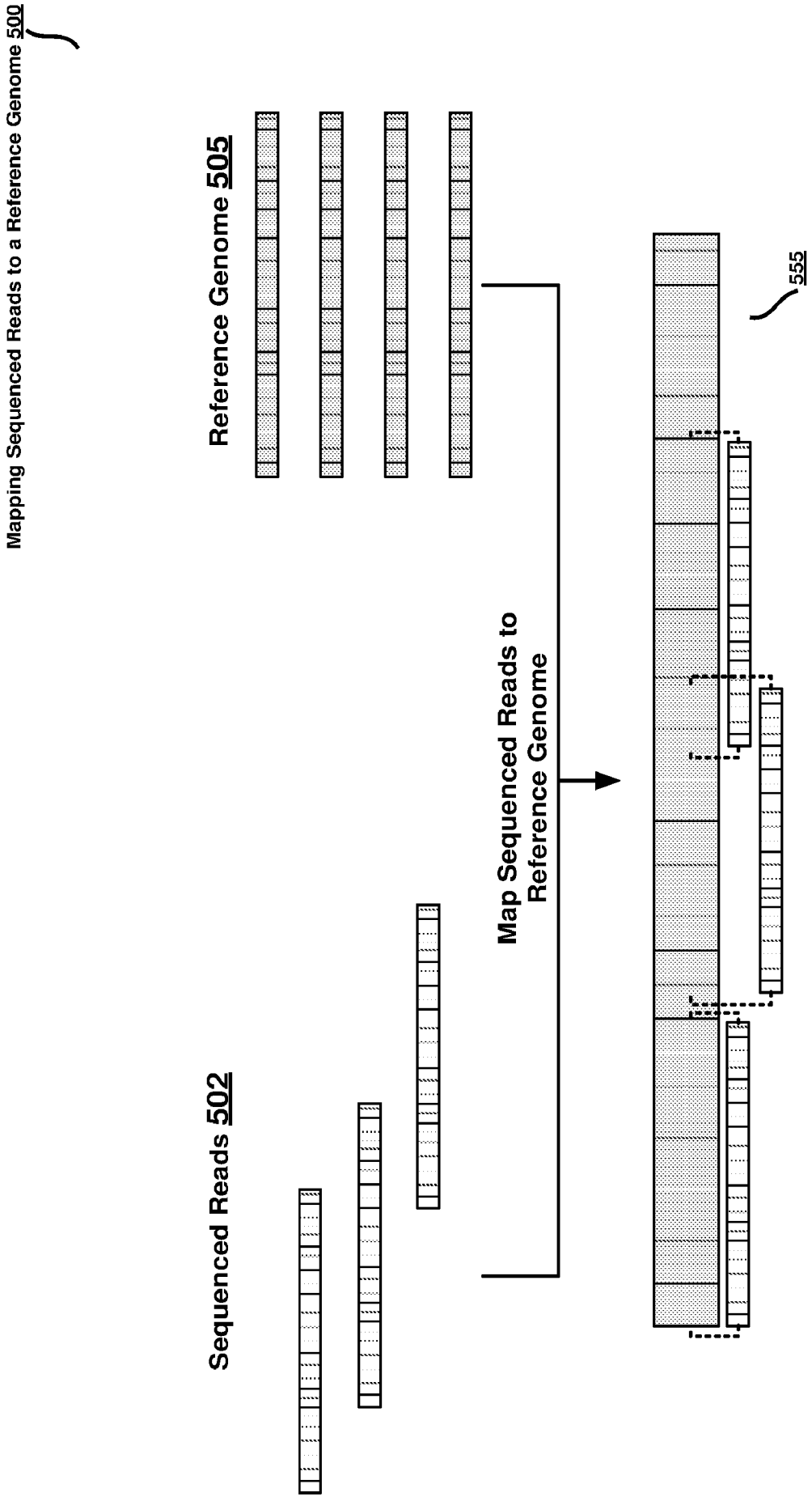

Mapping Sequenced Reads to a Reference Genome 500
Reference Genome 505
Sequenced Reads 502
Map Sequenced Reads to Reference Genome
555

Figure 5

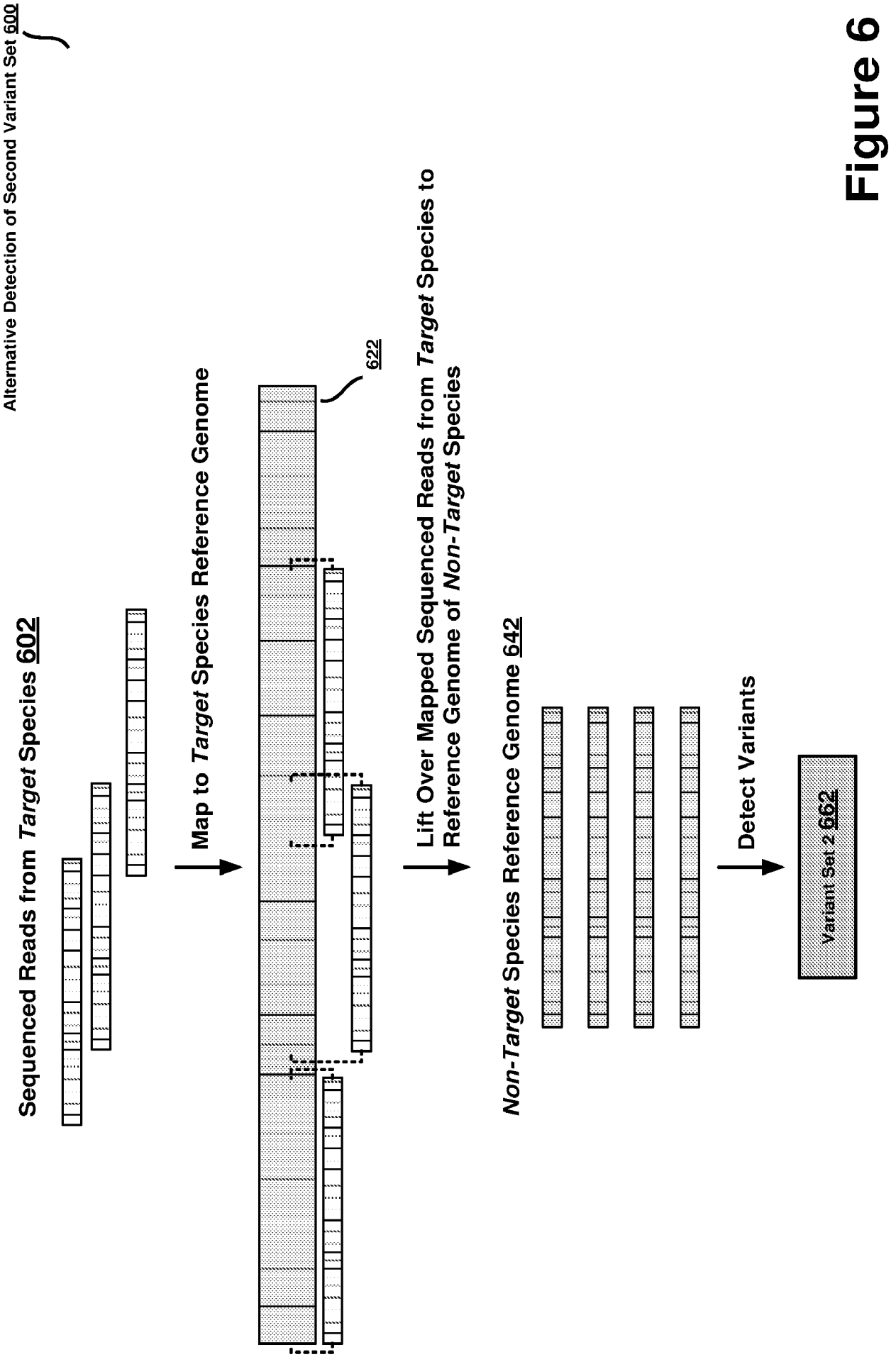

Alternative Detection of Second Variant Set 600
Sequenced Reads from *Target* Species 602
Map to *Target* Species Reference Genome
622
Lift Over Mapped Sequenced Reads from *Target* Species to Reference Genome of *Non-Target* Species
Non-Target Species Reference Genome 642
Detect Variants
Variant Set 2 662

Figure 6

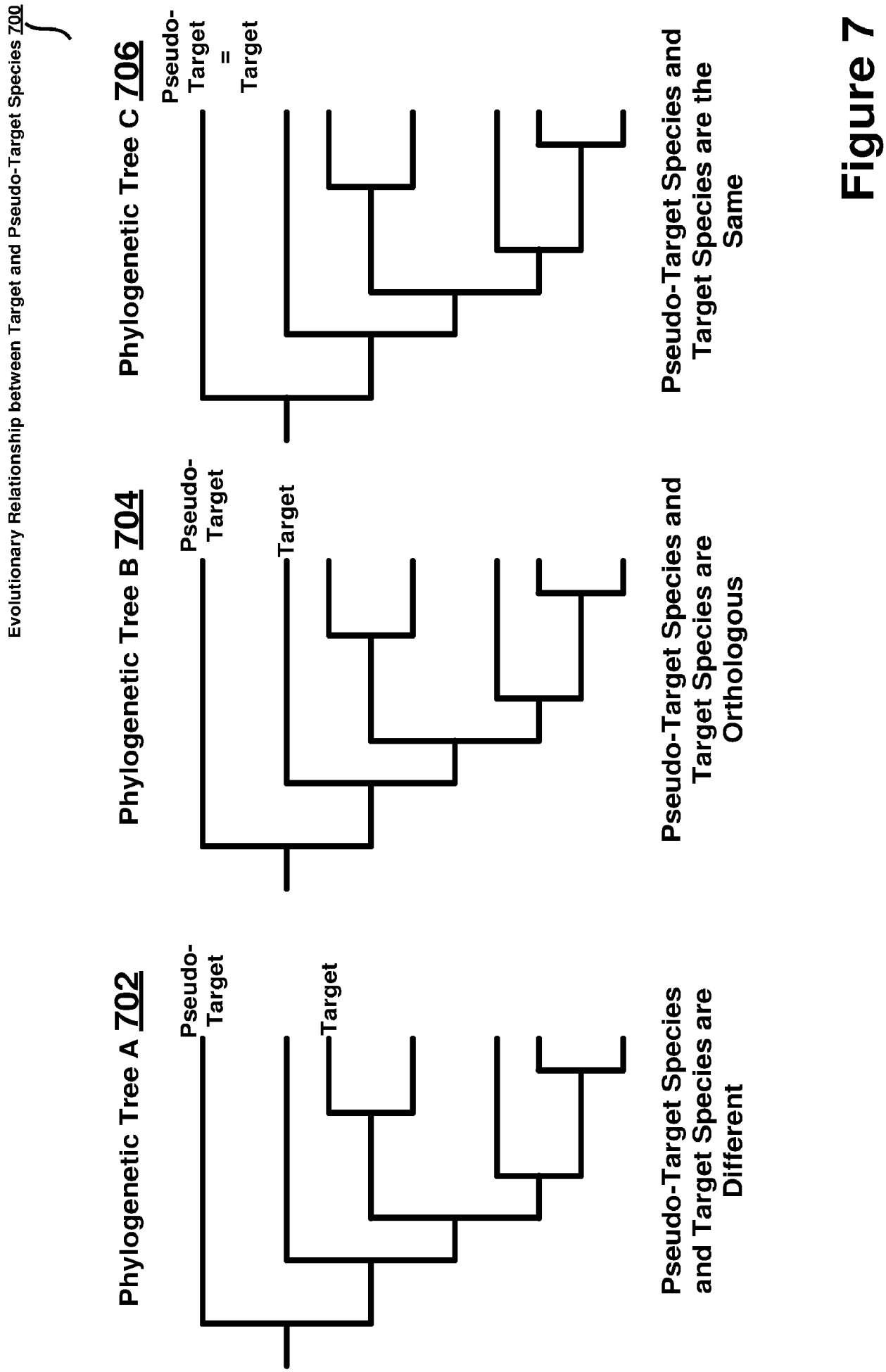

Evolutionary Relationship between Target and Pseudo-Target Species 700
Phylogenetic Tree A 702
Pseudo-Target
Target
Pseudo-Target Species and Target Species are Different
Phylogenetic Tree B 704
Pseudo-Target
Target
Pseudo-Target Species and Target Species are Orthologous
Phylogenetic Tree C 706
Pseudo-Target = Target
Pseudo-Target Species and Target Species are the Same

Figure 7

Evolutionary Relationship between Target and Non-Target Species 800

Phylogenetic Tree D 802

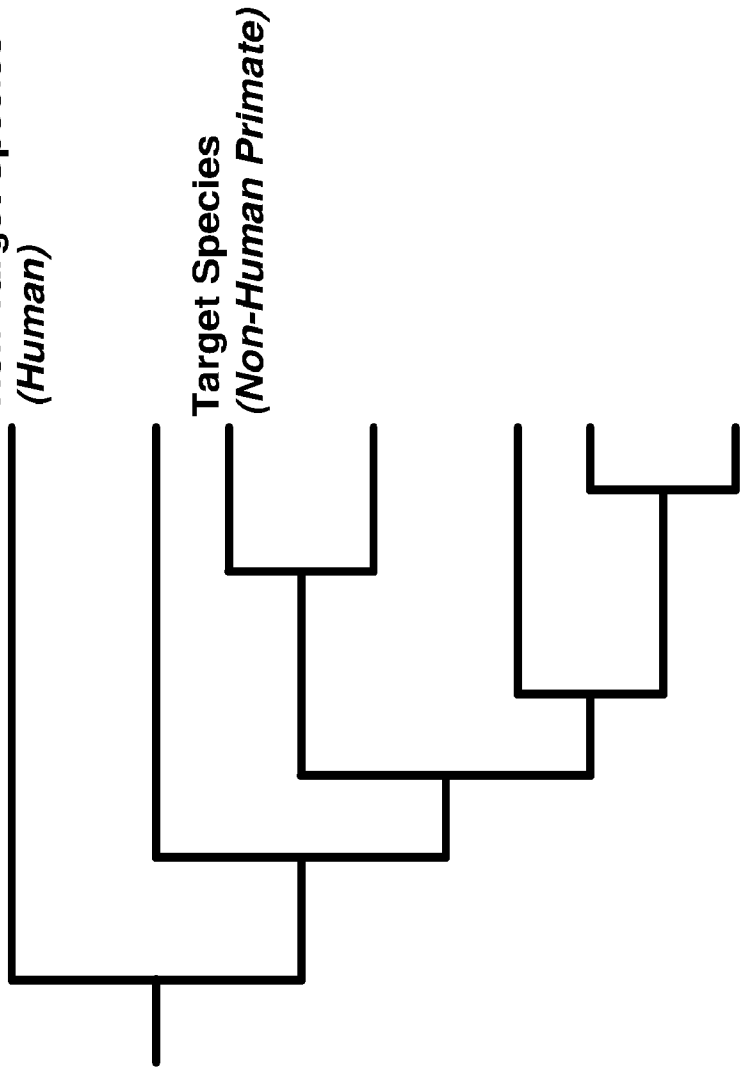


Figure 8

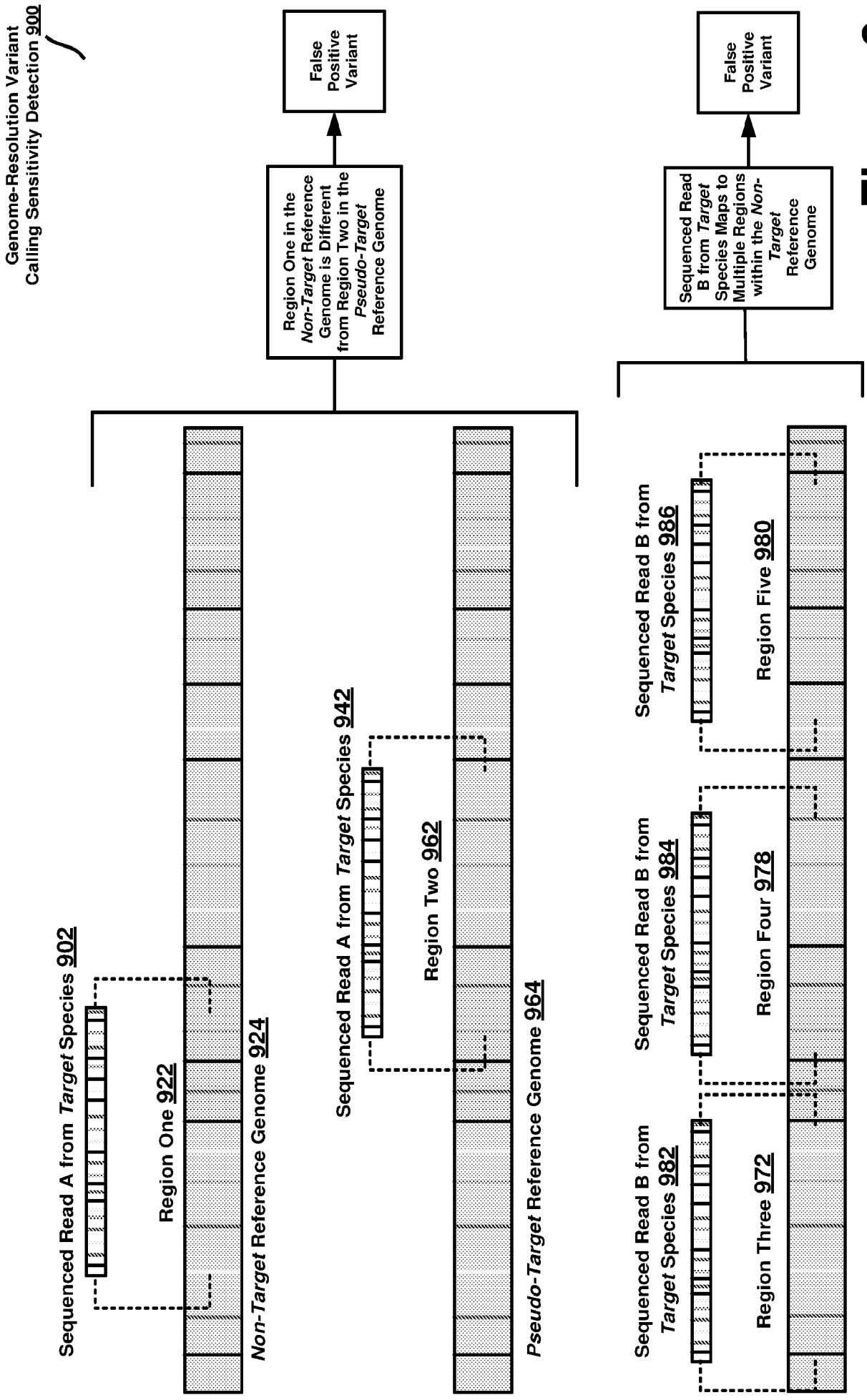

Genome-Resolution Variant Calling Sensitivity Detection 900
Sequenced Read A from Target Species 902
Region One 922
Non-Target Reference Genome 924
Sequenced Read A from Target Species 942
Region Two 962
Pseudo-Target Reference Genome 964
Region One in the Non-Target Reference Genome is Different from Region Two in the Pseudo-Target Reference Genome
False Positive Variant
Sequenced Read B from Target Species 982
Region Three 972
Sequenced Read B from Target Species 984
Region Four 978
Sequenced Read B from Target Species 986
Region Five 980
Sequenced Read B from Target Species Maps to Multiple Regions within the Non-Target Reference Genome
False Positive Variant

Figure 9

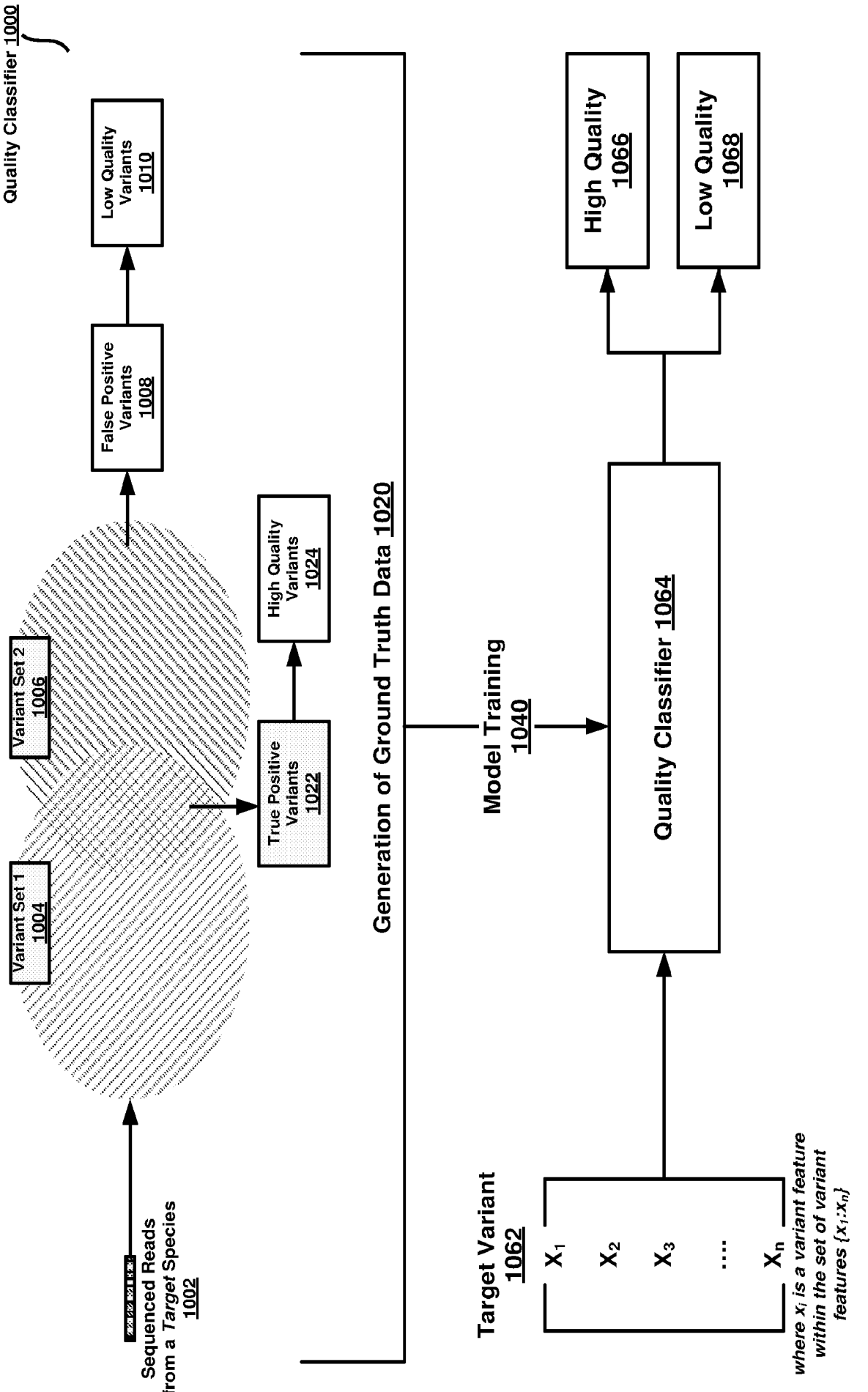

Quality Classifier 1000
Low Quality Variants 1010
False Positive Variants 1008
Variant Set 2 1006
Variant Set 1 1004
High Quality Variants 1024
True Positive Variants 1022
Sequenced Reads from a Target Species 1002
Generation of Ground Truth Data 1020
Model Training 1040
Target Variant 1062
X1
X2
X3
....
Xn
where xi is a variant feature within the set of variant features {x1:xn}
Quality Classifier 1064
High Quality 1066
Low Quality 1068

Figure 10

SUBSTITUTE SHEET (RULE 26)

Variant Features Describing Guanine-Cytosine Content 1100

Genetic Sequence B 1102

A	T	A	A	G	A	C	G	T	T	G	A	C	A	C

GC Content of Genetic Sequence B

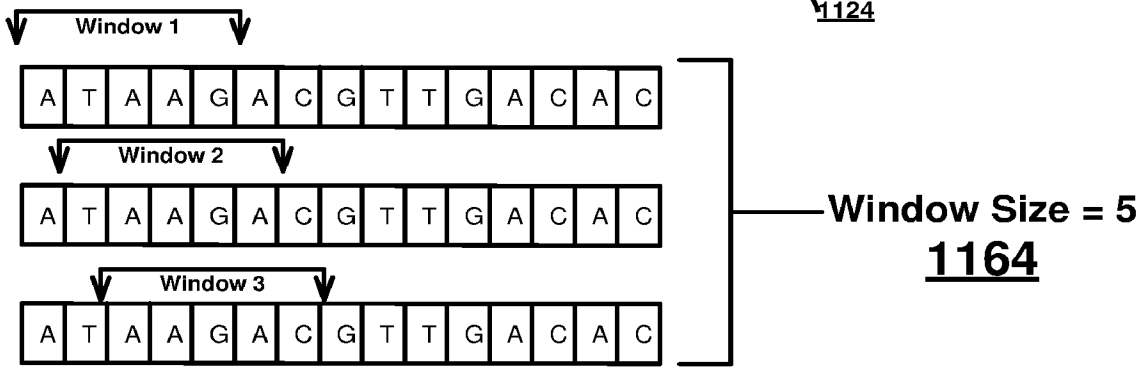
$$\frac{(G+C)}{(A+C+G+T)} = \frac{(3+3)}{(6+3+3+3)} = 0.4$$

1122

GC Skew of Genetic Sequence B

$$\frac{(G-C)}{(G+C)}$$

1124



Window	1	2	3	4	5	6	7	8	9	10	11
$\frac{(G-C)}{(G+C)}$	1/1	1/1	0/2	1/3	1/3	0/2	1/3	2/2	0/2	0/2	-1/3
Derivation from [G] = [C]	1	1	0	0.33	0.33	0	0.33	1	0	0	-0.33

Table 1184

Figure 11

SUBSTITUTE SHEET (RULE 26)

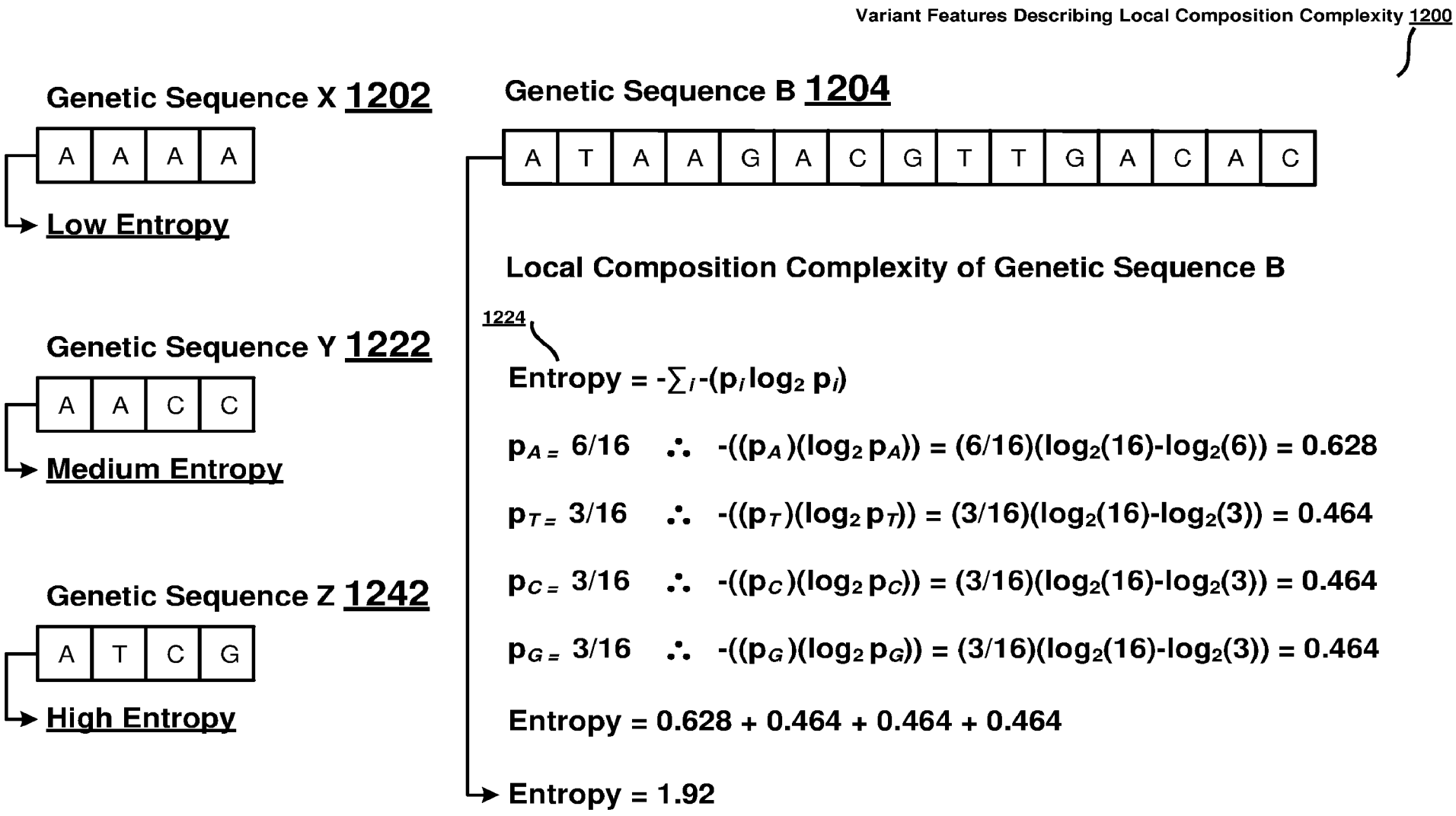


Figure 12

Variant Features Describing Allelic Count 1300
Variant 1 1302
A A C C
Variant 2 1304
A A T C
Population 1322
A A C C
A A T C
A A C C
A A T C
A A C C
A A T C
A A C C
A A T C
A A C C
A A T C
Allelic Count of Variant 1 = 6
Allelic Count of Variant 2 = 9

Figure 13

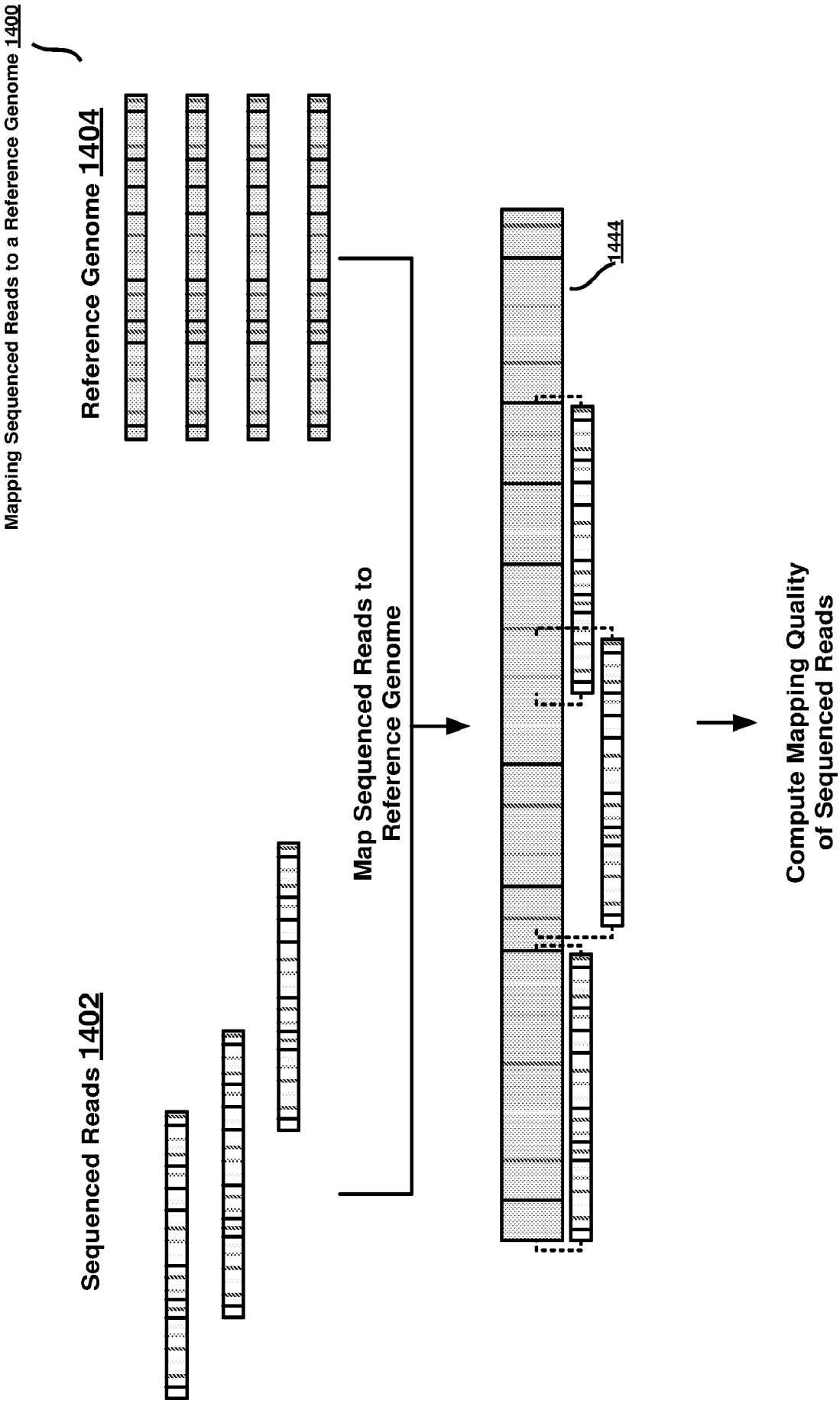

Mapping Sequenced Reads to a Reference Genome 1400
Reference Genome 1404
Sequenced Reads 1402
Map Sequenced Reads to Reference Genome
1444
Compute Mapping Quality of Sequenced Reads

Figure 14

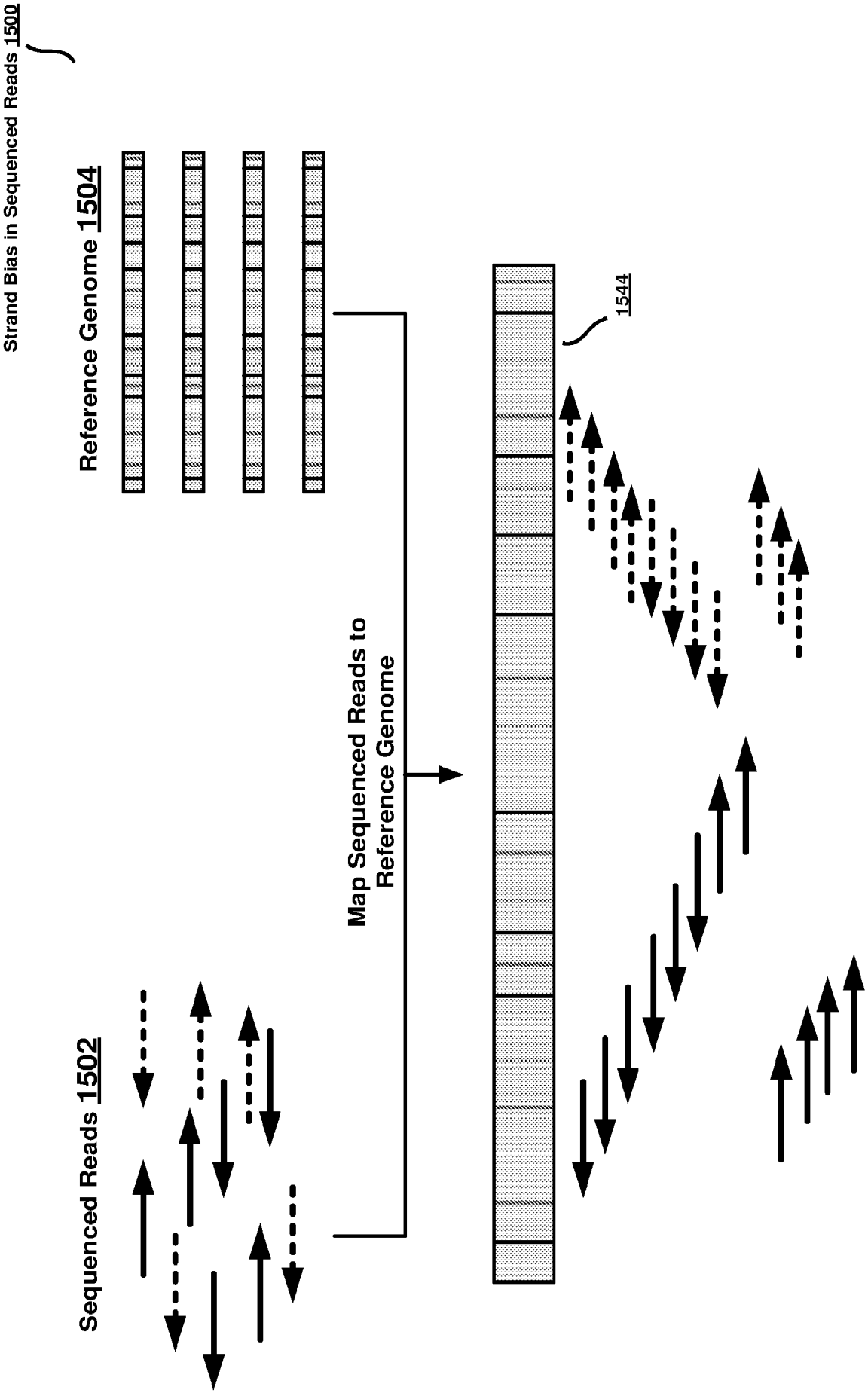

Strand Bias in Sequenced Reads 1500
Reference Genome 1504
Map Sequenced Reads to Reference Genome
1544
Sequenced Reads 1502

Figure 15

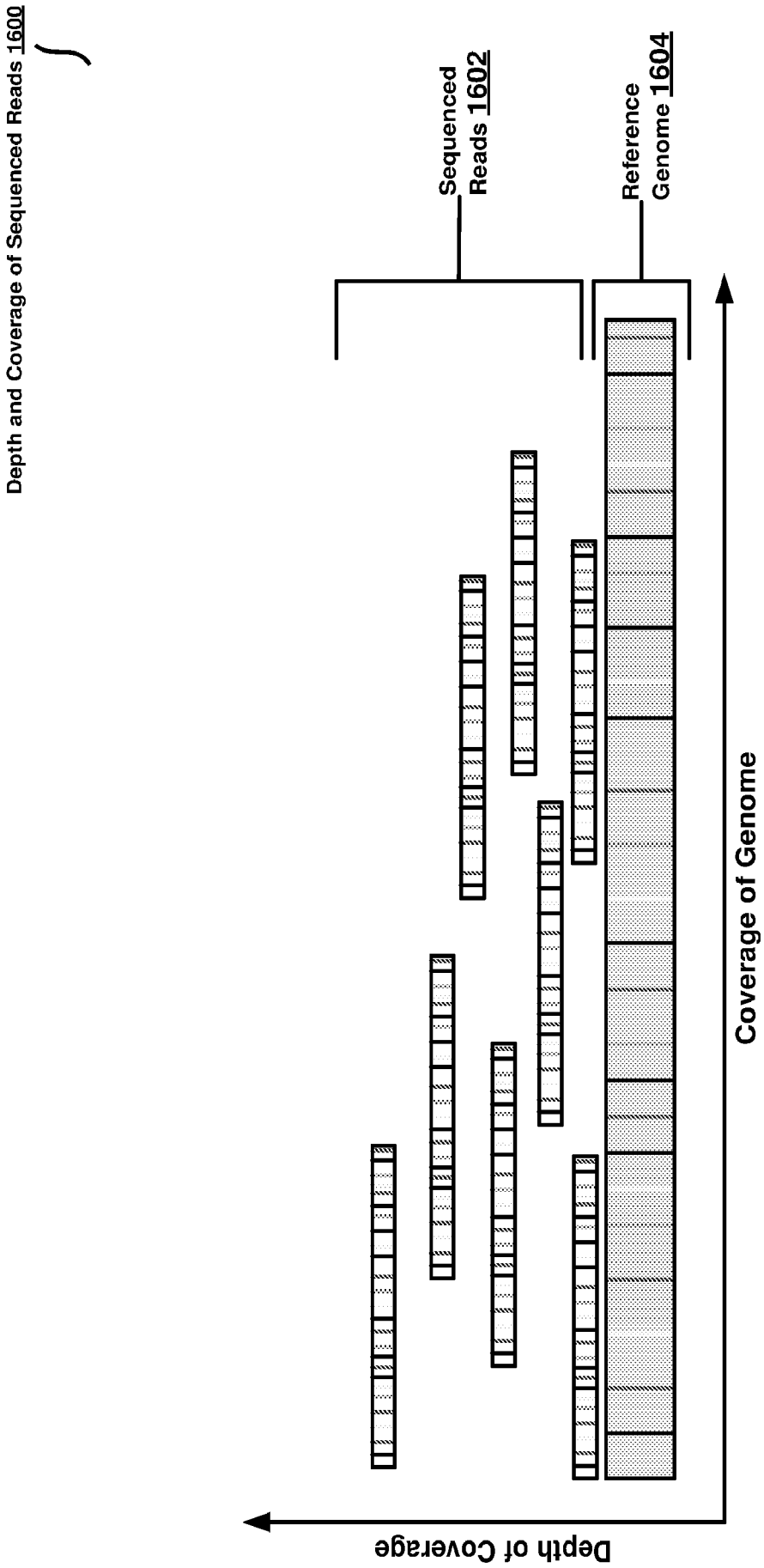

Depth and Coverage of Sequenced Reads 1600
Sequenced Reads 1602
Reference Genome 1604
Coverage of Genome
Depth of Coverage

Figure 16

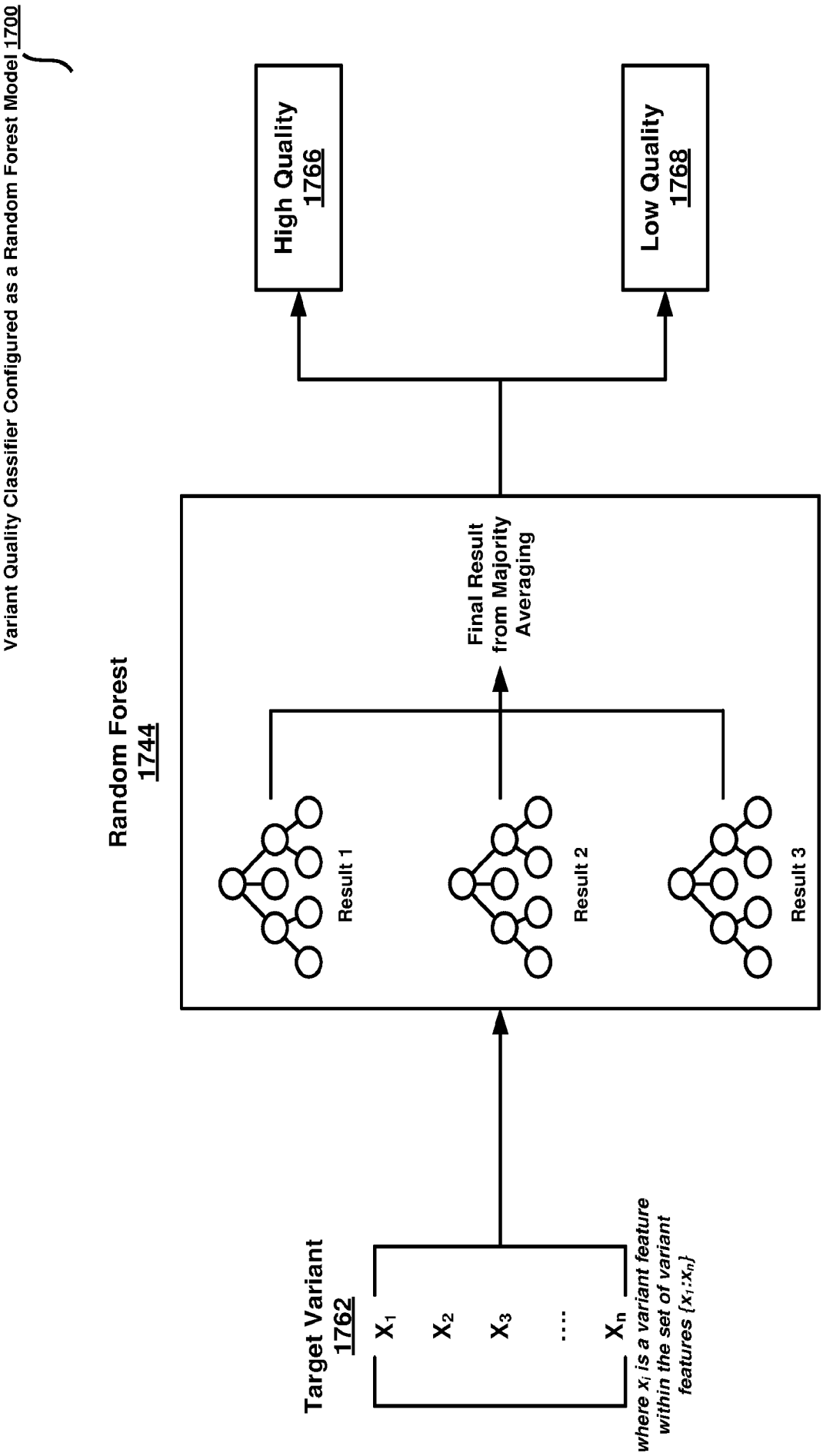

Variant Quality Classifier Configured as a Random Forest Model 1700
High Quality 1766
Low Quality 1768
Random Forest 1744
Final Result from Majority Averaging
Result 1
Result 2
Result 3
Target Variant 1762
X1
X2
X3
....
Xn
where xi is a variant feature within the set of variant features {x1:xn}

Figure 17

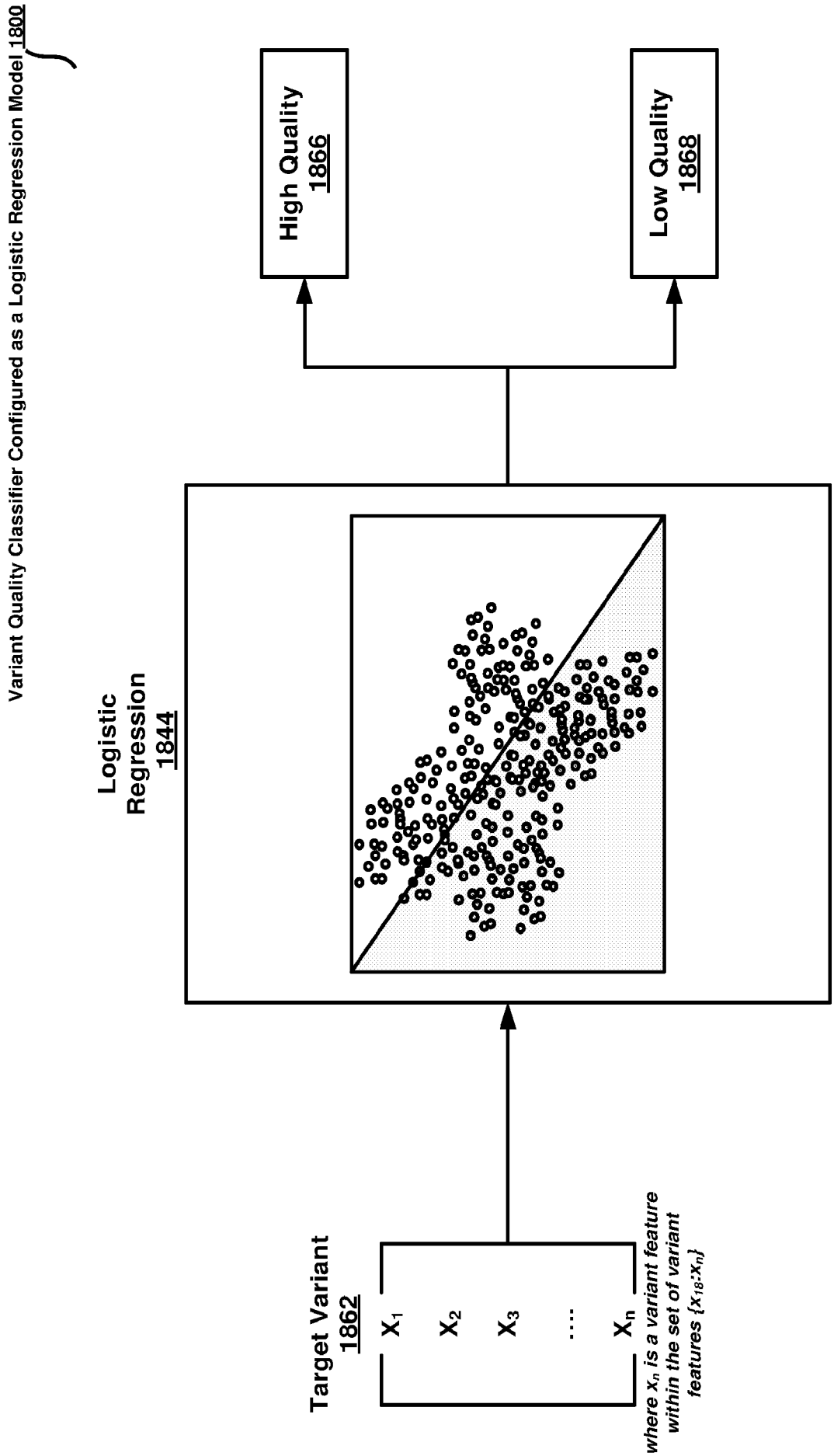

Variant Quality Classifier Configured as a Logistic Regression Model 1800
High Quality 1866
Low Quality 1868
Logistic Regression 1844
Target Variant 1862
X1
X2
X3
....
Xn
where xn is a variant feature within the set of variant features {x18:xn}

Figure 18

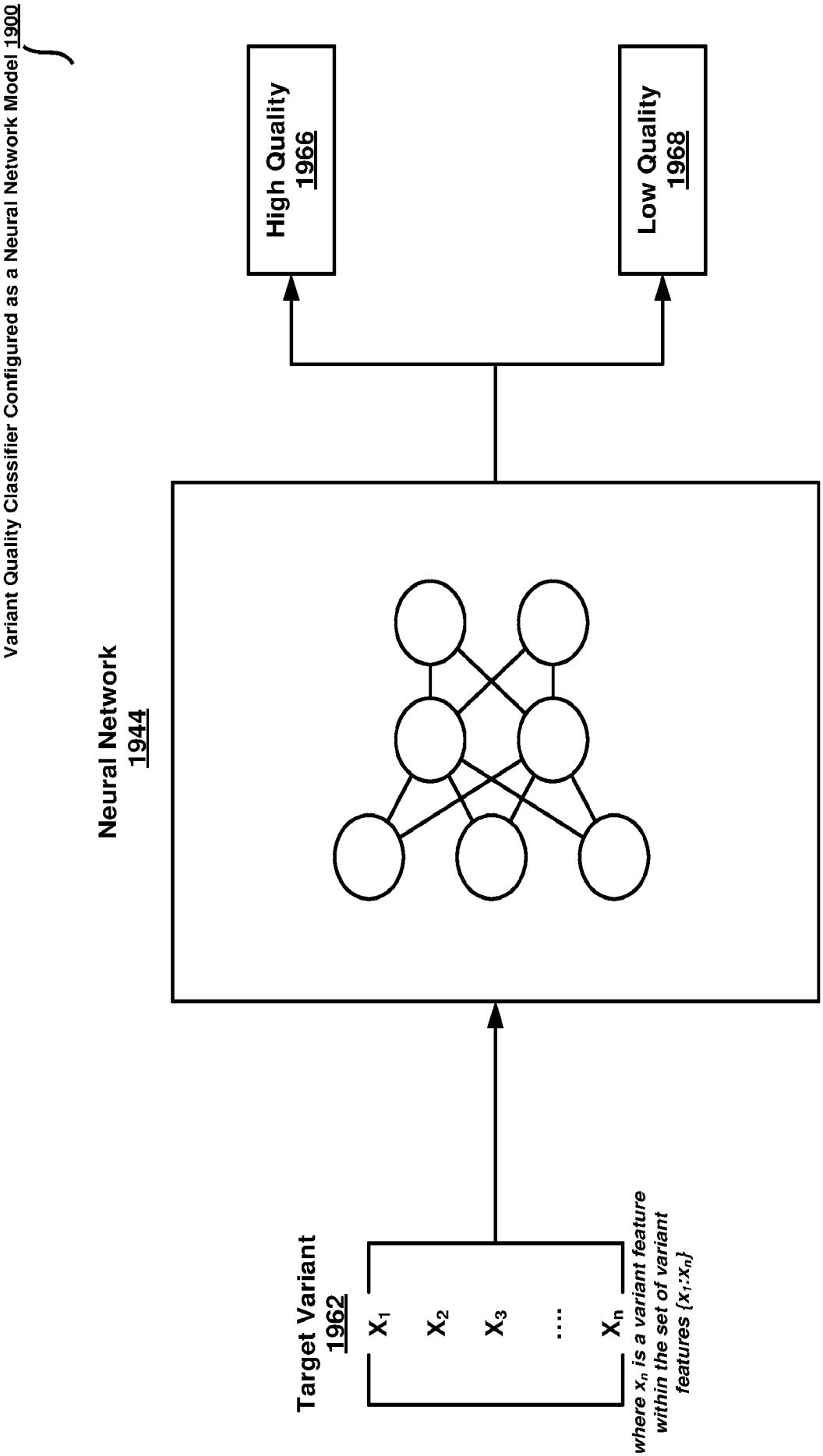

Variant Quality Classifier Configured as a Neural Network Model 1900
High Quality 1966
Low Quality 1968
Neural Network 1944
Target Variant 1962
X1
X2
X3
....
Xn
where xn is a variant feature within the set of variant features {x1:xn}

Figure 19

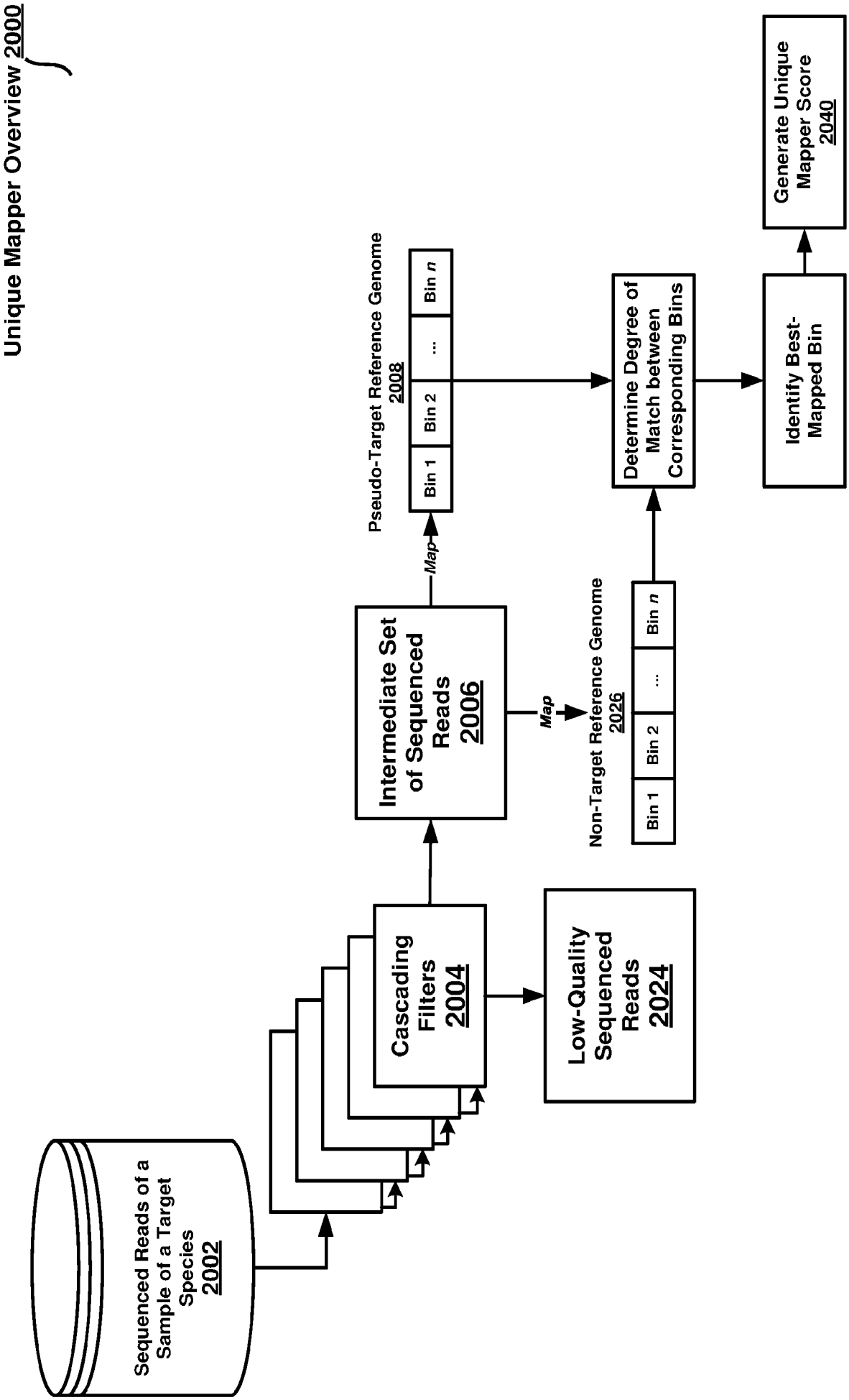

Unique Mapper Overview 2000
Sequenced Reads of a Sample of a Target Species 2002
Cascading Filters 2004
Low-Quality Sequenced Reads 2024
Intermediate Set of Sequenced Reads 2006
Map
Map
Pseudo-Target Reference Genome 2008
Bin 1
Bin 2
...
Bin n
Non-Target Reference Genome 2026
Bin 1
Bin 2
...
Bin n
Determine Degree of Match between Corresponding Bins
Identify Best-Mapped Bin
Generate Unique Mapper Score 2040

Figure 20

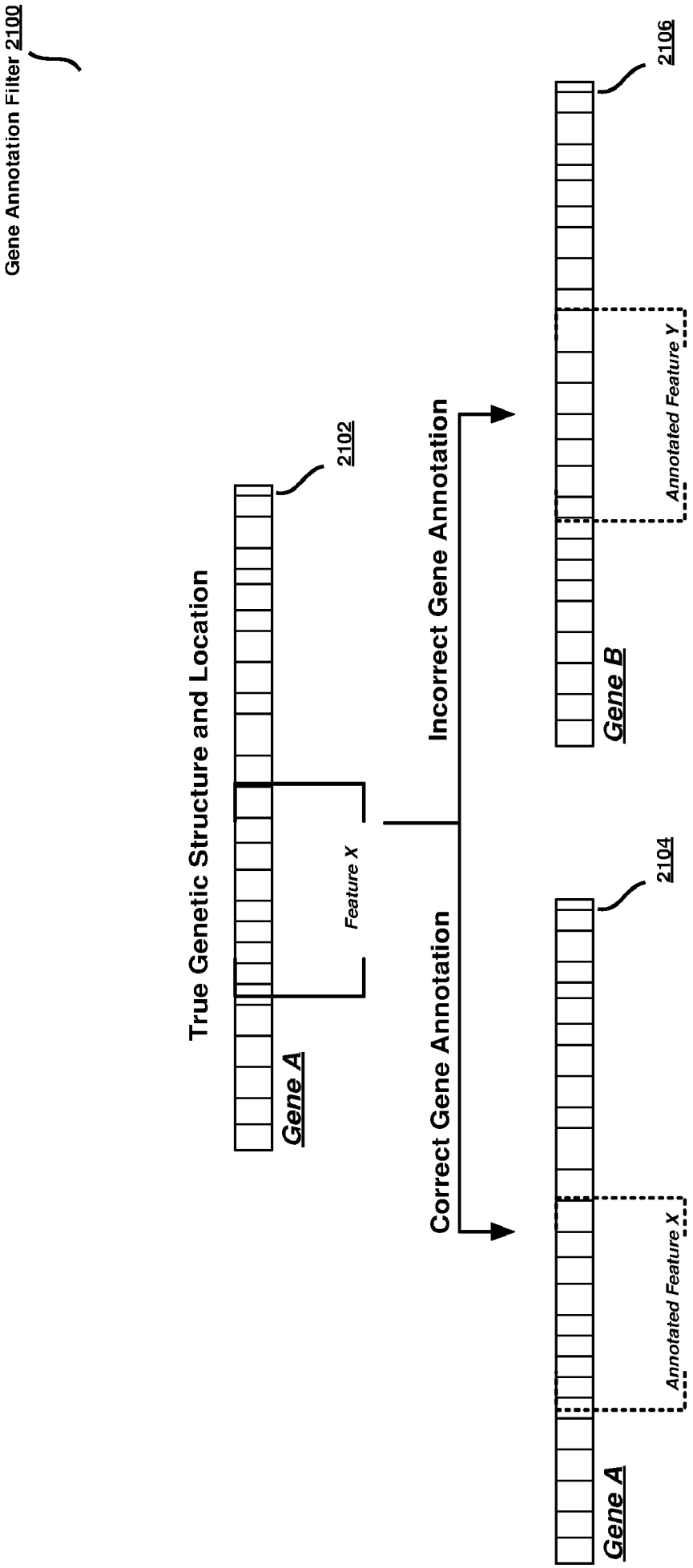

Gene Annotation Filter 2100
True Genetic Structure and Location
2102
Gene A
Feature X
Incorrect Gene Annotation
Correct Gene Annotation
2106
Annotated Feature Y
Gene B
2104
Annotated Feature X
Gene A

Figure 21

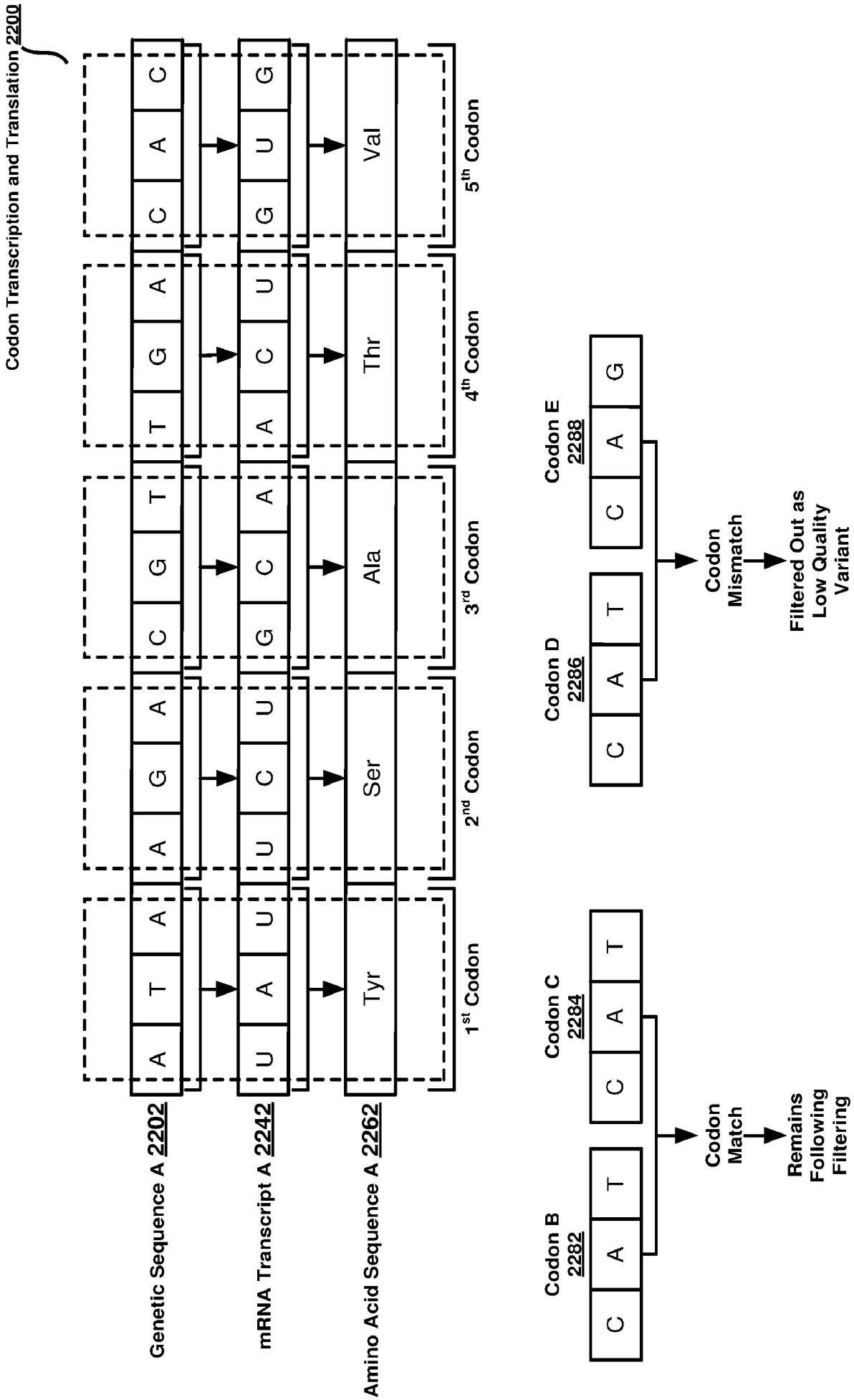

Codon Transcription and Translation 2200
Genetic Sequence A 2202
mRNA Transcript A 2242
Amino Acid Sequence A 2262
1st Codon
2nd Codon
3rd Codon
4th Codon
5th Codon
Tyr
Ser
Ala
Thr
Val
Codon B 2282
Codon C 2284
Codon D 2286
Codon E 2288
Codon Match
Remains Following Filtering
Codon Mismatch
Filtered Out as Low Quality Variant

Figure 22

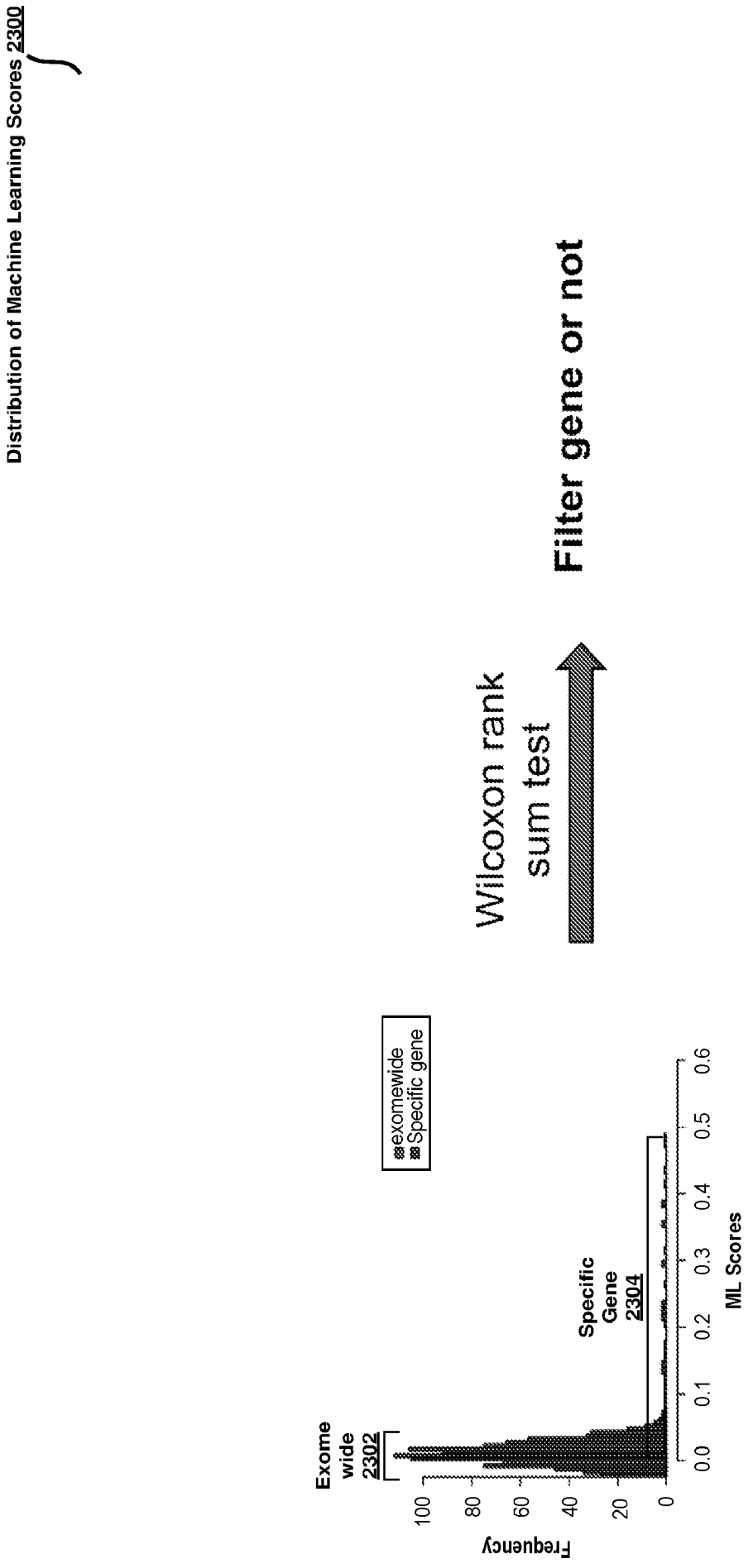

Distribution of Machine Learning Scores 2300
Filter gene or not
Wilcoxon rank sum test
exomewide
Specific gene
Exome wide 2302
Specific Gene 2304
Frequency
ML Scores
0
20
40
60
80
100
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 23

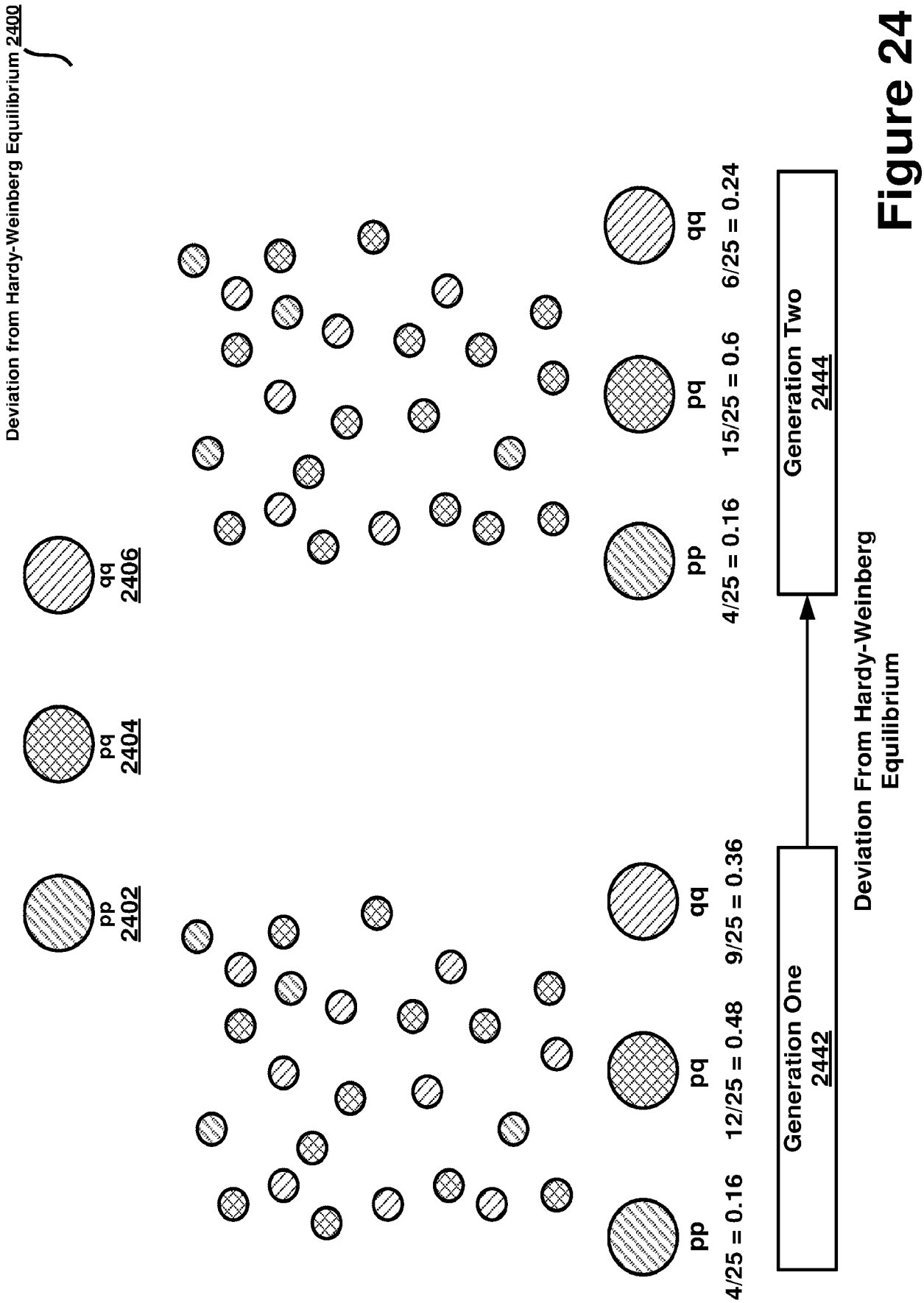

Deviation from Hardy-Weinberg Equilibrium 2400
pp 2402
pq 2404
qq 2406
Generation One 2442
4/25 = 0.16
12/25 = 0.48
9/25 = 0.36
Deviation From Hardy-Weinberg Equilibrium
Generation Two 2444
4/25 = 0.16
15/25 = 0.6
6/25 = 0.24

Figure 24

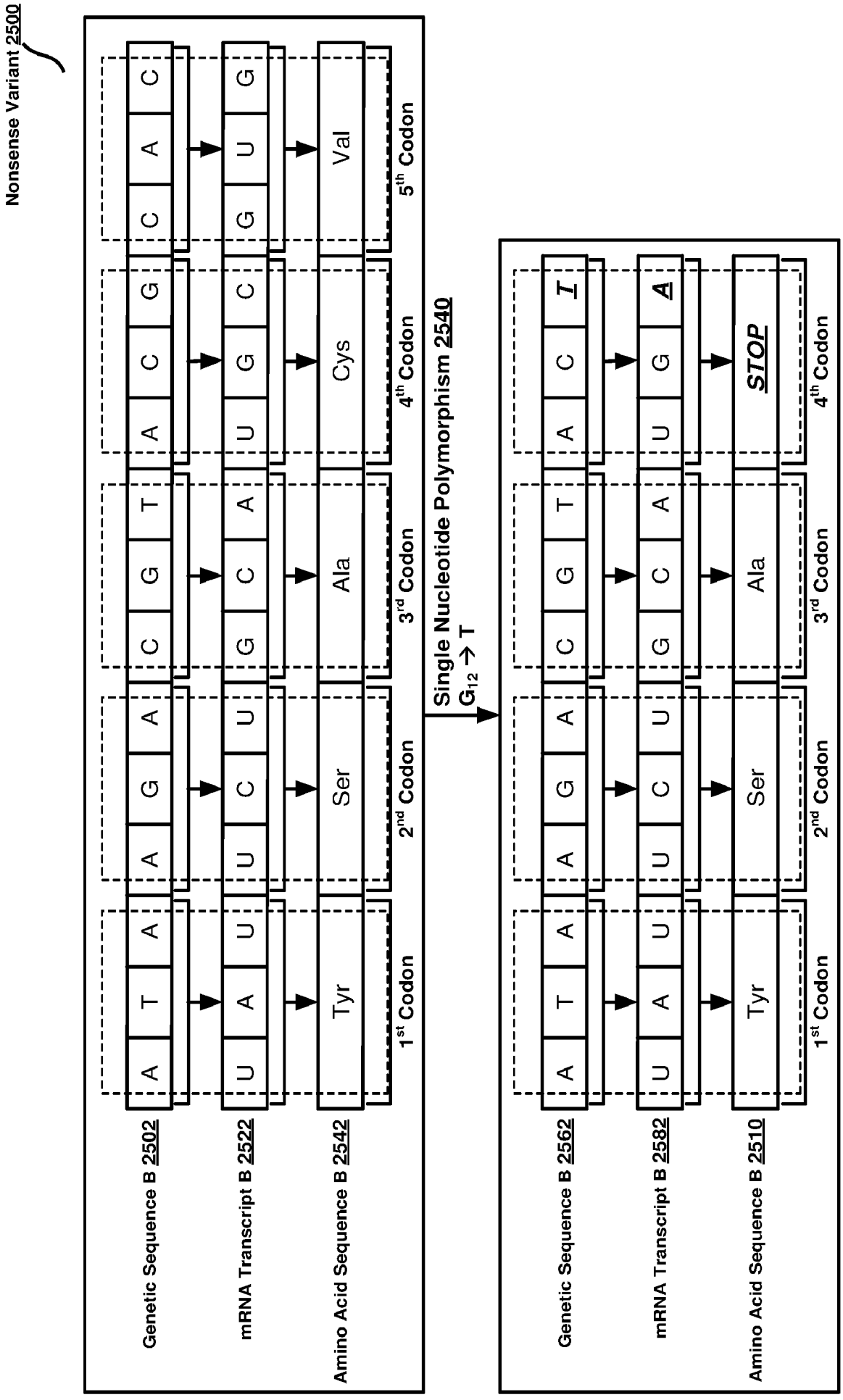

Nonsense Variant 2500
Genetic Sequence B 2502
mRNA Transcript B 2522
Amino Acid Sequence B 2542
1st Codon
2nd Codon
3rd Codon
4th Codon
5th Codon
Tyr
Ser
Ala
Cys
Val
Single Nucleotide Polymorphism 2540
G12 → T
Genetic Sequence B 2562
mRNA Transcript B 2582
Amino Acid Sequence B 2510
STOP

Figure 25

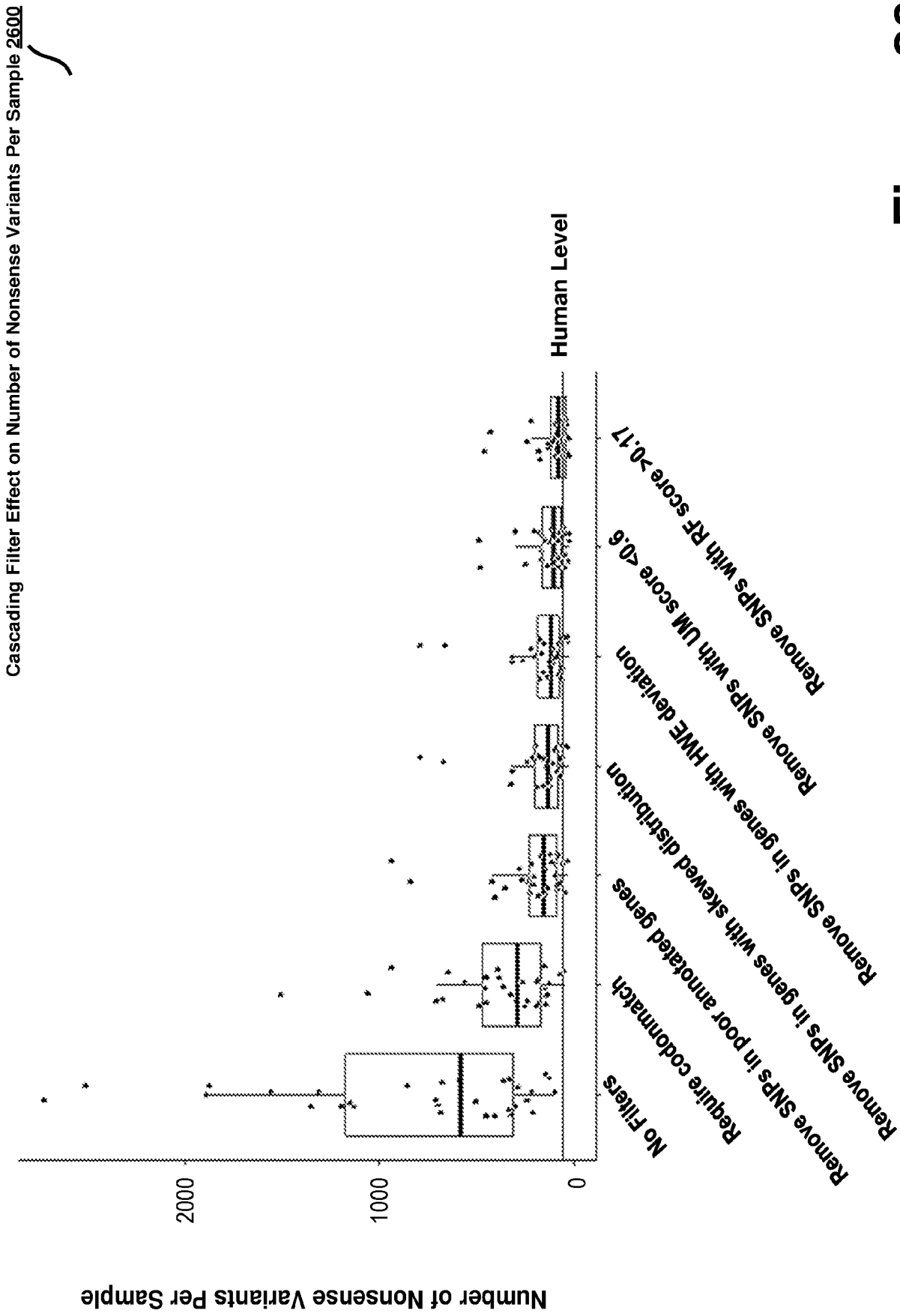

Cascading Filter Effect on Number of Nonsense Variants Per Sample 2600
Human Level
No Filters
Require codommatch
Remove SNPs in poor annotated genes
Remove SNPs in genes with skewed distribution
Remove SNPs in genes with HWE deviation
Remove SNPs with UM score <0.6
Remove SNPs with RF score >0.17
2000
1000
0
Number of Nonsense Variants Per Sample

Figure 26

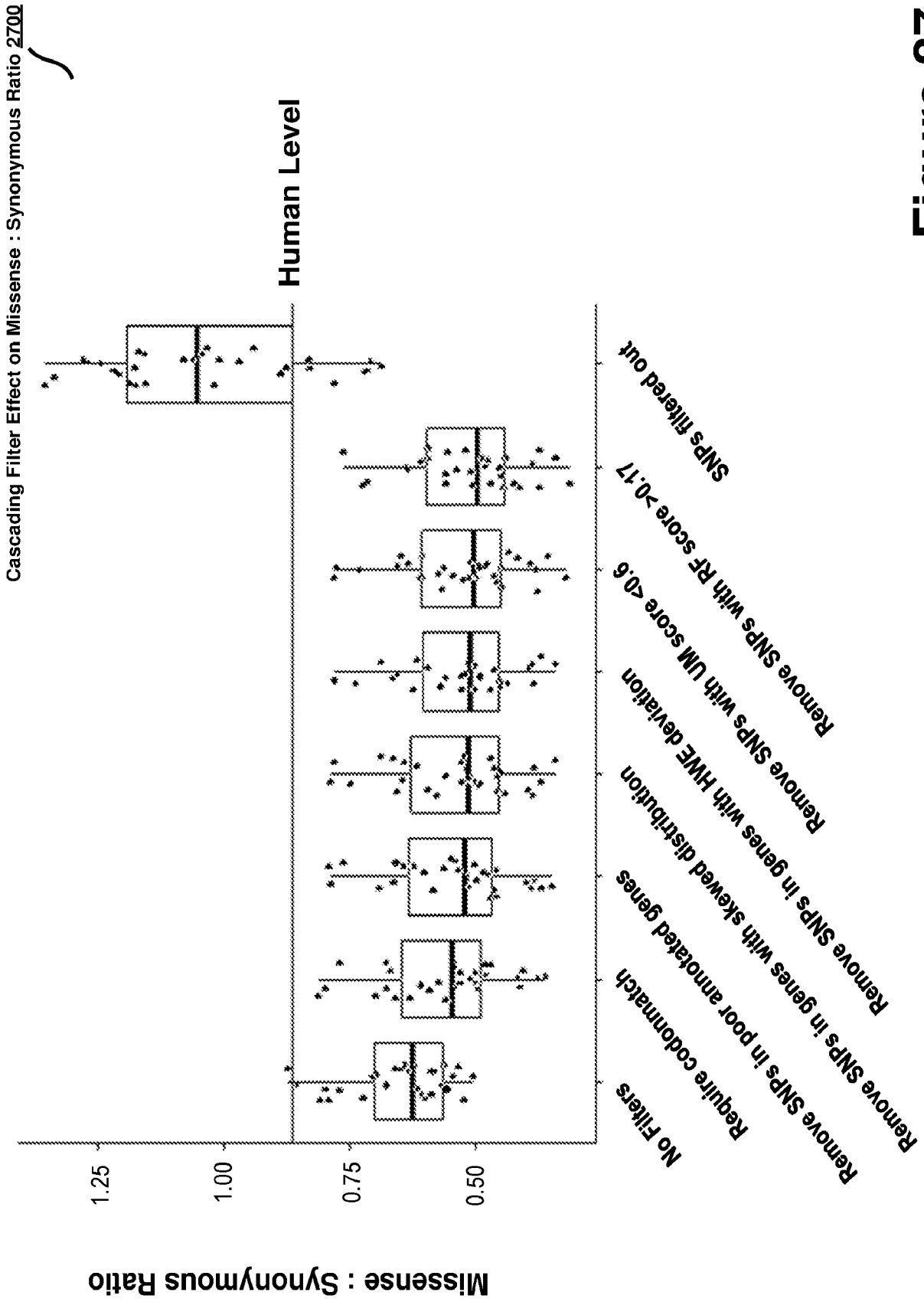

Cascading Filter Effect on Missense : Synonymous Ratio 2700
Human Level
Missense : Synonymous Ratio
1.25
1.00
0.75
0.50
No Filters
Require codommatch
Remove SNPs in poor annotated genes
Remove SNPs in genes with skewed distribution
Remove SNPs in genes with HWE deviation
Remove SNPs with UM score <0.6
Remove SNPs with RF score >0.17
SNPs filtered out

Figure 27

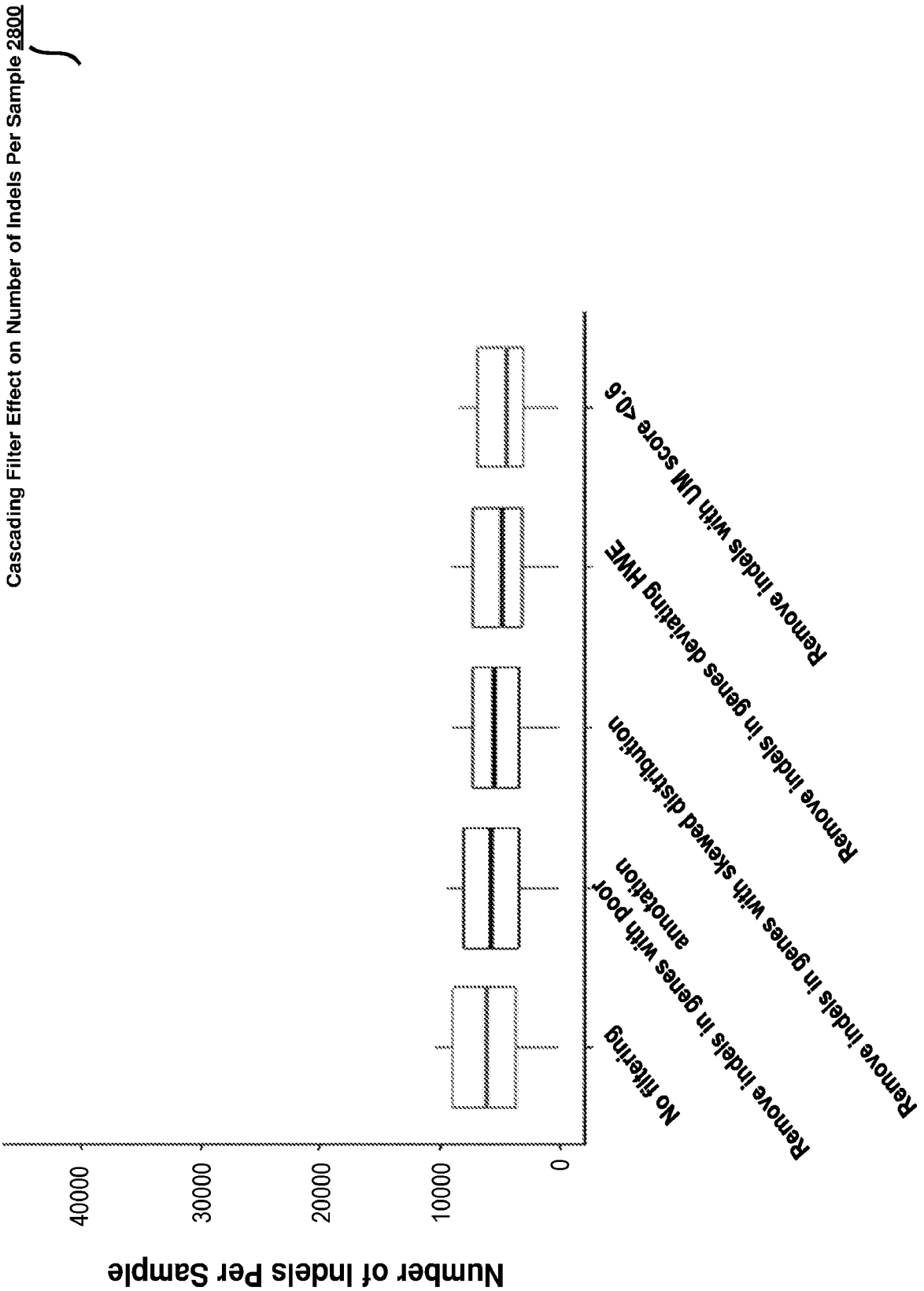

Cascading Filter Effect on Number of Indels Per Sample 2800
Number of Indels Per Sample
40000
30000
20000
10000
0
No filtering
Remove indels in genes with poor annotation
Remove indels in genes with skewed distribution
Remove indels in genes deviating HWE
Remove indels with UM score <0.6

Figure 28

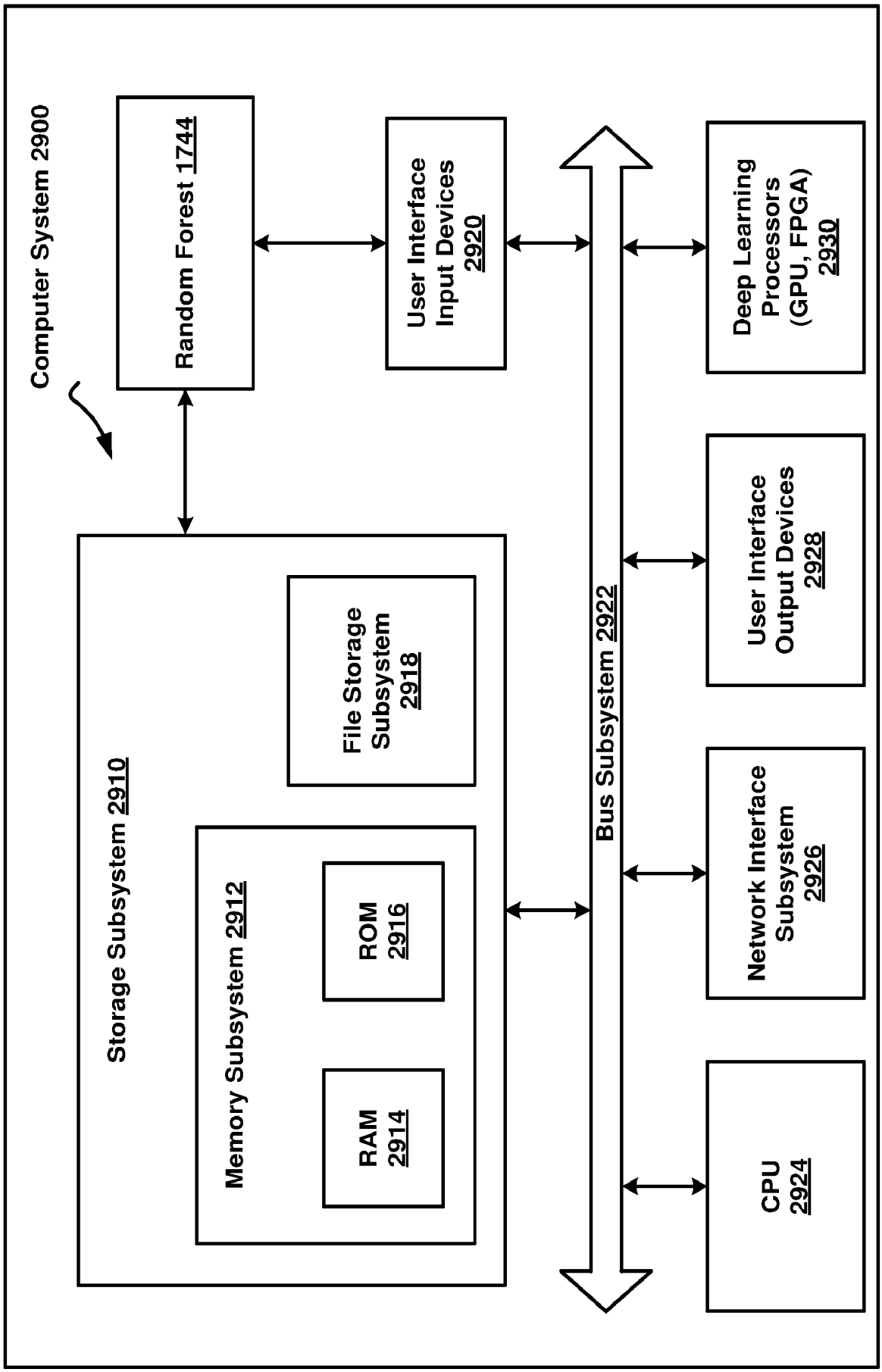

Computer System 2900
Random Forest 1744
User Interface Input Devices 2920
Deep Learning Processors (GPU, FPGA) 2930
Storage Subsystem 2910
File Storage Subsystem 2918
Memory Subsystem 2912
ROM 2916
RAM 2914
Bus Subsystem 2922
User Interface Output Devices 2928
Network Interface Subsystem 2926
CPU 2924

Figure 29