Machine learning-based detection of insertions and deletions in the human genome

Charles Curnin[1,2*], Rachel L. Goldfeder[3*], Shruti Marwaha[1,2], Devon Bonner[2], Daryl Waggott[1,2], Undiagnosed Diseases Network, Matthew T. Wheeler[1,2], Euan A. Ashley[1,2,4]

1. Division of Cardiovascular Medicine, Stanford School of Medicine.
2. Stanford Center for Undiagnosed Diseases.
3. The Jackson Laboratory for Genomic Medicine.
4. Department of Genetics, Stanford School of Medicine.

* These authors contributed equally.

Correspondence
Euan A. Ashley
Falk CVRB
870 Quarry Road
Stanford, 94305
euan@stanford.edu

Abstract

Insertions and deletions (indels) make a critical but under-studied contribution to human genetic variation. While indel calling has improved significantly, it lags dramatically in performance relative to single-nucleotide variant calling, something of particular concern for clinical genomics where larger scale disruption of the open reading frame can commonly cause disease. Here, we present a machine learning-based approach to the detection of indel breakpoints. Our novel approach improves sensitivity to larger variants by leveraging sequencing metrics and signatures of poor read alignment. We use new benchmark datasets and Sanger sequencing to compare our approach to current gold standard indel callers, achieving unprecedented levels of precision and recall. We demonstrate the impact of these calling improvements by applying this tool to a cohort of patients with undiagnosed disease, generating plausible candidates in 21 out of 26 cases. We highlight the diagnosis of one patient with a 498-bp deletion in *HNRNPA1* missed by traditional indel-detection tools.

Introduction

Insertions and deletions within the genome are well-established mechanisms of human disease[1]. While less common than single-nucleotide variants (proportions of incidence range from 1:7 to 1:43, varying highly with genomic region), indels are an important component of genetic diversity, and are more likely to disrupt the open reading frame[2,3]. In the latest version of the Human Genome Mutation Database (HGMD 2018.4) indel mutations account for 31% of all entries, with deletions outnumbering insertions more than 2:1[4].

While performance in identifying single-nucleotide variants (SNVs) is comparatively high, detecting indels with high sensitivity remains a challenge[5]. The current diagnosis rate through exome sequencing for patients with genetic disorders is around 30%[6,7]. By comparison with long-read sequencing, indel calling from short-read sequencing has been shown to miss variants, including clinically relevant ones[8]. Even with long-read sequencing, which remains relatively costly, indels are a persistent source of error[9]. One review of seven commonly used indel calling tools found that about 77% of indels across 78 human genomes from the 1000 Genomes Project went undetected by all tools[10]. Further complicating identification of real variants, there is often low concordance among these tools[11].

Evaluating the performance an indel caller involves comparing the variants it identifies against a "benchmark" or "truth" set of variants that we accept as fully characterizing the variation of the genome (possibly within certain genomic regions and/or categories of variation). This process produces scores of recall (sensitivity) and precision (positive predictive value), derived from the proportions of true positive, false negative, and false positive calls. These metrics capture how likely a genuine variant is to be called, and how likely a called variant is to be genuine. The ideal indel caller has high recall and high precision, reporting many genuine variants and few false ones; it moreover detects a variety of indels, including variants of different sizes, and those that lie in different types of genomic regions (e.g., homopolymer runs).

Benchmark data derived from finding consensus between multiple orthogonal variant callers is often labeled "gold standard." These variant callers may not individually detect all true positives in the genome; moreover, while the requirement of including only variants identified by multiple callers ensures that the variants in the benchmark set are of high confidence, it may leave out additional genuine variants. Incompleteness of benchmark datasets can warp

benchmarking metrics, flagging real calls made by callers outside the truth set as false positives (deflating estimates of precision). The decreased number of true positives may also inflate estimates of recall. Additionally, we note that machine-learning approaches trained on an incomplete benchmark dataset are limited, learning to classify real indels missing from the truth set as normal loci.

There are currently several categories of indel callers that detect different types of indel signatures. Split-read tools, such as SV-M[12], extract read pairs where only one mate can be confidently mapped to the reference, and align the unmapped mates. Paired-end tools, like BreakDancer[13], examine the distribution of insert sizes between mate reads in a given region to determine whether sequence has been added or removed. Alignment-based tools like Stampy[14] examine how reads align to a reference. Some alignment-based tools, for example Platypus[15], may conjecture alternate haplotypes to which reads are aligned. Others, such as Scalpel[16], perform graph-based alignments. For sensitive detection of larger indels, there are fewer choices. The most popular tool may be Pindel[17], which uses a pattern-growth algorithm to detect indels through split reads. Other tools, like IMSindel[18] and ScanIndel[19], use *de novo* assembly to identify large variants.

Smaller indels (1 to 5 bp) make up a majority (77%) of indels by number, but they account for a very small proportion (6%) of all inserted and deleted bases in the human genome. By number of nucleotides, larger indels exert significantly more influence on the diversity of the genome. But, as indel size increases, the performance of most tools declines. In the narrow range of sizes of 1 to 10 bp, Hasan et al. found that the F-score (the harmonic mean of recall and precision) of seven popular indel-detection packages drops by nearly half[10]. Some callers appear axiomatically restricted in the size of indels they report. Applied to simulated variants of up to 50 bp in length, popular callers such as GATK UnifiedGenotyper[20], SAMTools[21], and VarScan[22] detected no indels greater than 37, 44, and 42 bp in length, respectively[23].

A distinct category of tools exists for detecting copy-number variants (CNVs) and structural variant (SVs), large-scale genetic abnormalities of a kilobase or more in length. But this leaves few options for sensitive detection of indels larger than a few bases and smaller than one kilobase. Insertions, which cannot be detected through changes in sequencing depth, can be particularly challenging. Our goal was to build on the capabilities of recent tools and leverage the availability of improved benchmark datasets to develop an indel caller with increased sensitivity to variants across the size spectrum.

Results

A benchmark genome of simulated indels facilitates evaluation

One source of benchmark data is simulated variants. The primary advantage presented by this technique is that the truth set is known with maximal confidence, since variants are precisely "spiked in." This improves the reliability of calling metrics. Additionally, the incidence of different kinds of indels can be manipulated to generate sufficient test data to characterize indel callers' performance on a variety of variants. The primary drawback of simulated indels is that they may not perfectly recapitulate real genotypes.

There exist several tools that can be precisely directed to spike variants into a starting genome. Using BAMSurgeon[24], we defined 3,885 non-overlapping indels, three with each possible size from 5 to 1000 bp—and then one indel for each 10 bp increment between 1000 and

10000 bp, on average. Half were deletions, and half were insertions for which the inserted sequence was generated randomly. The positions for these variants were selected randomly from mappable regions of chromosome 22 at least 10 bp away from known pre-existing indels in a starting genome (NA12878). When benchmarking variant callers' performance on this dataset, we also excluded the known pre-existing true positives in NA12878 and calls made by callers within 10 bp of them. We note the benchmark-incompleteness issue still exists here: variants in NA12878 but missing from its truth set will be present in this dataset, and callers' capture of them may incorrectly be flagged as false positives, deflating estimates of precision.

A minority of these variants (n = 860) were rejected by BAMSurgeon because they could not be made to appear biologically realistic, usually because starting coverage was too low, or a sufficiently large contig into which to inject the variants could not be assembled. Above 8,000 bp in length, we found, it was increasingly unlikely that the indels could be introduced. We used this dataset—which is distinct from the simulated variants used to supplement the training data of our new tool—to benchmark the variant callers.

Syndip offers a comprehensive set of variants for benchmarking

"Gold standard" datasets available for benchmarking include public genomes such as NA12878[25] and Syndip[26]. NA12878 is the genome of a woman from Utah, whose variants are characterized by the Genome in a Bottle Consortium (GiaB) according to consensus calling across multiple sequencing and variant calling platforms. Syndip is a recently released synthetic diploid genome produced by combining two haploid human cell lines sequenced using single molecule real-time sequencing and identifying indels with FermiKit[27], FreeBayes[28], Platypus[15], Samtools[29], GATK HaplotypeCaller and GATK UnifiedGenotyper[30] (Fig. 1).

Relative to NA12878, Syndip offers a wider range of indels that provide for more comprehensive benchmarking. Syndip includes more indels than NA12878, and these variants span a greater range in size. While across NA12878, the mean and standard deviation of indel sizes is just 3 and 4 bp, in Syndip it is 22 and 209 bp. And while in NA12878, the largest variant is just 127 bp, in Syndip, it is 19 kb. Syndip was also developed with long-read sequencing, which incurs random errors that can generally be overcome by sequencing depth. Neither of the two machine learning-based callers evaluated here (DeepVariant and Scotch) were trained on Syndip. GATK HaplotypeCaller may have an advantage as it was one of the original tools used to develop the Syndip truth set.

Both NA12878 and Syndip aim to describe human genomes, which should be nearly identical in terms of the number and range of sizes of indels detected. It is highly unlikely that Syndip contains significantly more indels, or significantly larger ones; instead, we believe the discrepancy in variant composition suggest that tools used to analyze NA12878 were likely less sensitive to larger variants. Both datasets also contain fewer indels as size increases toward their edge of detection capabilities. This may be a function of the attenuating performance of the chosen variants callers. In an analysis of short-read whole-genome sequencing data from 2,504 individuals, the 1000 Genomes Project reported that—above 50 bp—the number of deletions of a given size varies little with that size—up to roughly 10 kb. Above 50 bp, moreover, the number of insertions of a given size, generally increases with that size, with a peak around 9 kb[31].

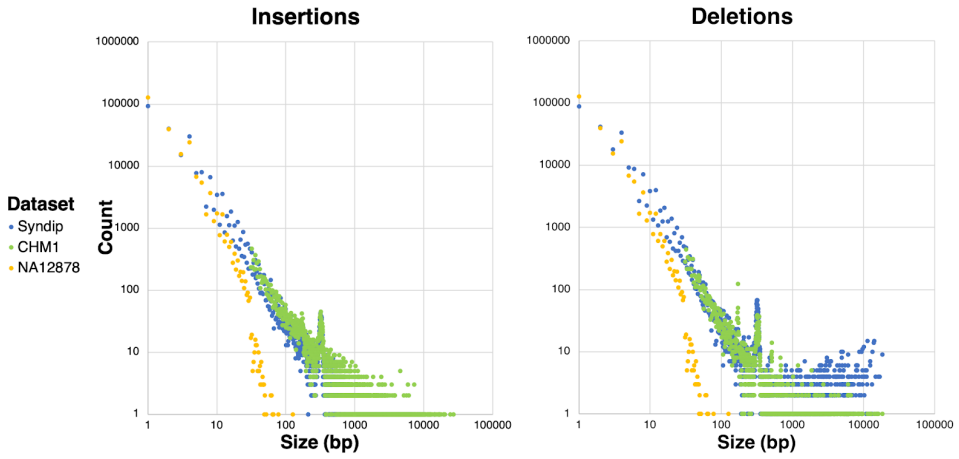

Fig. 1: Log-scale distribution of indels by size by dataset.
Despite both deriving from human genomes, NA12878 contains fewer indels than Syndip, and smaller ones. We also examine here CHM1[32], a haploid complete hydatidiform mole for which there exists a comprehensive list of indels larger than 30 bp detected by SMRT sequencing. Ostensibly, Syndip, which consists of two cell lines (CHM1 and CHM13), should contain more indels than CHM1 alone. But even within certain size ranges, CHM1 has more variants. While both datasets have a peak in the number of indels around 320 bp (representing SINEs, short interspersed nuclear elements), CHM1 alone has an additional peak around 170 bp[33]. These variants are either false positives in CHM1, or false negatives in Syndip.

A machine-learning based caller designed for capturing large indels

We present a machine learning-based tool focused on increasing the sensitivity of calling larger indels from whole-genome sequencing data (Scotch, Fig. 2). Machine learning techniques have been applied to variant calling with success before. DeepVariant[34], which uses neural networks to analyze pileups, won highest "SNP Performance" in the precisionFDA Truth Challenge. Scotch examines designated portions of the genome, and analyzes each base individually. It creates a numerical profile of these positions, describing various features of the aligned sequencing data like depth, base quality, and alignment to the reference. A full explanation of the features selected is available in the Supplementary Note.

Based on these predictors, a random forest model then classifies the position as non-indel or the site of a specific type of indel breakpoint. If the locus does not match the reference, Scotch classifies it as one of three types of indel breakpoints—the site of an insertion, the start of a deletion, or the end of a deletion—or as a 1-bp deletion, which requires a separate class since both deletion breakpoints fall on the same locus, considered in half-open notation. The Scotch standard model was trained on the NA12878 genome. We added to its training data larger simulated indels as a way of attempting to overcome the incompleteness of the benchmark dataset, which generally lacks larger indels.

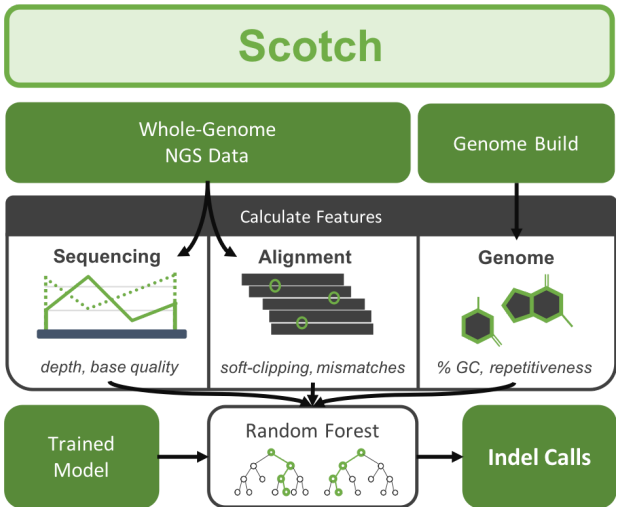

Fig. 2: Scotch is a machine-learning based indel caller.
Features are calculated from input sequencing data and from a reference genome. A random forest identifies positions that are the breakpoints of an insertion or deletion.

We evaluated Scotch and five other callers: DeepVariant; GATK HaplotypeCaller[30]; VarScan2[22]; and two versions of Pindel, the standard, which we refer to as "Pindel", and the pipeline with the "-l" option for reporting long insertions, "Pindel-L". These versions of Pindel call the same deletions; the "-l" version includes many additional insertions. We assess the performance of these six pipelines on three benchmark datasets: simulated variants, Syndip[26], and NA12878[25]. For ease of use, we subset chromosome 22 from each of these datasets. The full results of this benchmarking are available in Supplementary Tables 1 - 9. Below, we concentrate primarily on the simulated variant dataset that contains many large variants, and Syndip, which offers the most comprehensive set of variants.

Protocol for indel benchmarking

Given a truth set for a benchmark genome and the query set of variants reported by an indel caller in the same, we perform a breakpoint-based comparison to generate metrics describing the caller's performance. We subdivide each list into separate registers of insertion, deletion start, and deletion end breakpoints. Corresponding truth-query pairs are input to an app developed by the Global Alliance for Genomics and Health (GA4GH Benchmarking) made available on precisionFDA[35]. In a distance-based comparison, it categorizes calls as true positives, false positives, or false negatives, and uses this information to derive precision and recall scores. While not explicitly a part of the comparison, this method does credit or penalize tools for their estimate of the size of a deletion, which determines where the caller places its start and end breakpoints. It does not rely on alternate allele sequences, which Scotch, Pindel, and Pindel-L do not report for all variants.

This process produces recall and precision scores for each class of breakpoint. For deletions, we report the mean metric across start and end breakpoints. For each tool, we also benchmarked the full set of all called indel breakpoints against all true positive breakpoints. This illustrates a caller's performance on both insertions and deletions, with an emphasis on performance on deletions, since in aligned sequencing data each deletion comprises two breakpoints (start, end) while an insertion has only one. Combining all breakpoints also

expresses a tool's ability to determine that an indel of some sort exists at a given locus, even if the type is not correctly identified.

We separate recall into two metrics: recall by count, which considers the numbers of variants identified; and recall by base, which considers the numbers of bases belonging to the variants identified. The latter highlights a caller's sensitivity to larger variants. GA4GH Benchmarking reports recall by count directly; we calculate recall by base by considering the number of bases belonging to true positive and false negative variant calls.

Optimizing for high recall

In general, we chose to value recall over precision. While low precision is addressed in clinical pipelines through filtering steps and manual curation that eliminate out false positives, low recall means losing variants that cannot be recovered. Estimates of precision, moreover, can mislead if the caller identifies true positives that are missing from the benchmark set. Even on simulated data, a tool may be penalized for identifying an undisclosed indel in the starting genome on which a variant simulation tool operates. Below, we motivate this rationale further through Sanger sequencing (which reveals that many calls flagged as false positive are in fact genuine variants).

Scotch has high recall on simulated data, identifying indels of up to several kilobases

Evaluated on the dataset of simulated indels, most callers perform well in identifying small variants. However, their performance generally declines markedly as indel size increases (Fig. 3). In contrast, across all indel breakpoints, Scotch's recall by count and recall by base (99%; 99%) both exceed Pindel-L (79%; 74%), which itself surpasses all other tools. On deletions, Scotch (recall by count: 97.9%), is incrementally superior to Pindel (97.2%), and far exceeds other callers (which range in recall from 1% to 12%). On insertions, the differences are even more clear: Scotch (recall by count: 98%) surpasses Pindel-L (42%) and all other tools (0.3% - 24%). Scotch retains high recall across the size spectrum, successfully identifying insertions and deletions in the dataset larger than the largest variant on which it was trained (500 bp), and including the dataset's largest insertion of 7810 bp and largest deletion of 7608 bp.

Across this broad range of indel sizes, we note that the full relationship between a caller's recall and indel size is complex. While, in general, as indel size increases, sensitivity decreases, there are important exceptions. Pindel-L registers a sharp decline, then increase in recall around 1 kb. For most callers, the steepest drop in recall occurs near 150 bp, approximately the length of a short read. These variants may be particularly difficult to detect, because unlike smaller variants, they cannot be contained in a single short read. Beyond 150 bp, recall is still somewhat variable, and some callers continue to decline in performance while others rebound slightly.

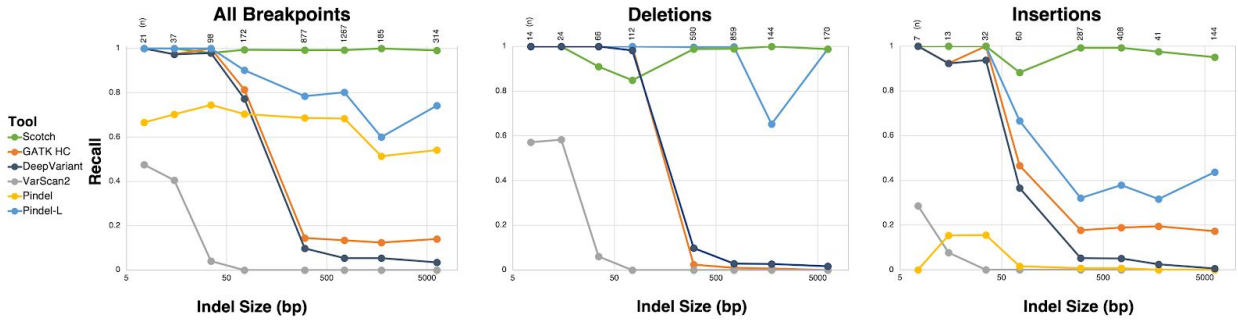

Fig. 3: Recall by indel size on simulated variants.
To examine how callers' performance varies with indel size, we group together variants with sizes within a fixed range and assess the caller's sensitivity to the group. For each group, represented here as a data point, we indicate above the number of variants in the group. The performance of many popular callers declines as indel size increases. Scotch's performance, in contrast, is more consistent across the size spectrum. (Note: Since Pindel and Pindel-L call the same deletions, their recall curves on these variants are overlapping.) This data is also available in Supplementary Tables 13 - 15.

Scotch identifies variants in Syndip with high recall
Similar performance is seen when these tools are tested on Syndip (Supplementary Tables 1 - 3). Across all indel breakpoints, Scotch's recall by count is the highest of any tool (93%), exceeding Pindel-L (91%), DeepVariant (87%), GATK HaplotypeCaller (87%), VarScan2 (73%), and Pindel (66%). The variants identified by DeepVariant, GATK HaplotypeCaller, and VarScan2 account for, between 37% and 14% of all inserted and deleted sequence.

Scotch has low precision on consensus benchmark data
While Scotch has high recall, on consensus data sets, it exhibits low precision. By precision, Scotch (28%), Pindel (27%) and Pindel-L (6%), fall far behind VarScan2 (98%), DeepVariant (95%), and GATK HaplotypeCaller (91%). Estimates of Scotch's precision, however, may be deflated by its identification of real variants missing from the truth set. As discussed below, Sanger sequencing of Scotch's calls flagged in benchmarking as false positives reveals that many are *bona fide* variants. Note also the differential metrics: Scotch's insertion-specific precision is 13%, while its deletion-specific precision is 72%.

Metal: a meta-analytic indel caller sensitive to large variants
Each of the benchmarked callers has its own strengths, and none outperforms all others in all circumstances. To produce an optimal variant calling tool, we merge their strengths[36].

Some tools achieve excellent performance in one dimension by sacrificing performance in another. VarScan2, for example, is very conservative, and thus attains extremely high precision by accepting low recall. Here, instead, we attempt to negotiate a "best compromise." This approach, which we call Metal, retains the high precision of DeepVariant, GATK HaplotypeCaller, and VarScan2, while incorporating many of the larger variants that Scotch and Pindel-L identify. It achieves this by performing a "smart intersection." Metal will report a call produced by a tool if it has a corresponding call within 3 bp identified by another tool. To counter the low insertion-specific precision of Scotch and Pindel-L, we require that insertions called by these tools have correlates in higher-precision DeepVariant, GATK HaplotypeCaller, or VarScan2. Metal does not consider the various quality scores that tools report with the

variants they call, which may not be directly comparable because tools use different scales, but integrating this evidence is a promising area for further improvement. An additional machine learning model, in fact, could arbitrate the calls made by each tool and collate them into a single set with maximal confidence.

Across all indel breakpoints in Syndip, Metal's recall by count (91%) surpasses all tools except Scotch (93%), while tripling precision (85%, cf. Scotch: 28%, Fig. 4). Metal identifies far more large variants than DeepVariant, GATK HaplotypeCaller, or VarScan2, with a recall by base of 57%. On deletions, the performance benefits are particularly clear. Metal identifies more variants by count than any individual tool (90%), and more by base (60%) than all tools except Pindel and Pindel-L, with greatly improved precision (81%).

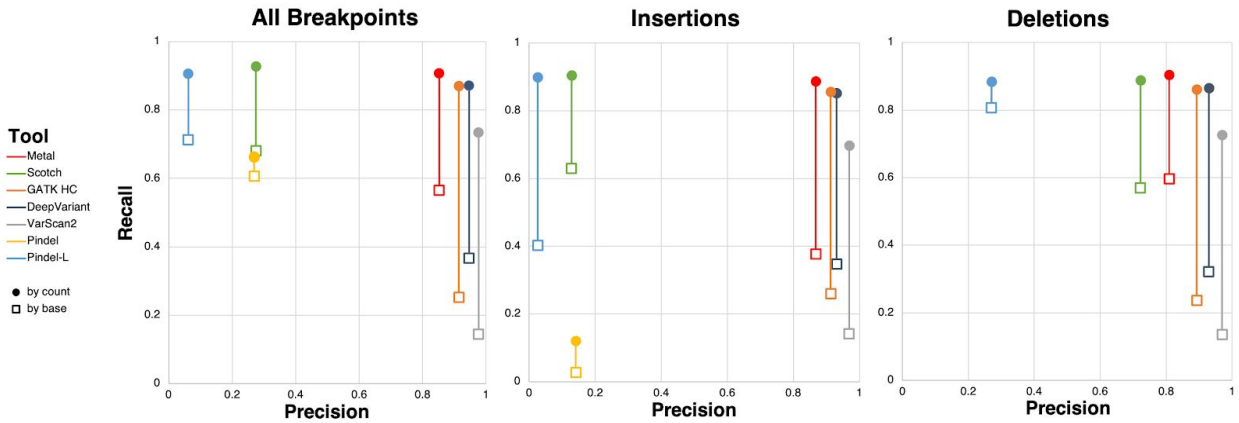

Fig. 4: Scotch and Metal offer high sensitivity to large variants and improved precision in Syndip. Performance of the selected indel callers, including Scotch and the meta caller Metal, on Syndip. Recent callers such as DeepVariant and GATK HaplotypeCaller have married high precision and high recall by count, but they are more likely to miss larger variants. Scotch and Pindel-L offer higher recall, especially on a per-base basis, but with lower precision. Metal, a combination of the other six pipelines (Scotch, GATK Haplotypecaller, DeepVariant, VarScan2, and the Pindel versions), captures some of the large variants Pindel-L and Scotch identify while sacrificing little in precision.

F(n) metrics balance recall and precision for clinical variant calling

An F-score balances considers recall, by base or by count, and precision. An F1-score computes the harmonic mean of recall and precision, giving each equal weight. In an F(N)-score, recall is considered N times more times more important than precision:

$$F(N) = (1 + N^2) \times \frac{precision \times recall}{(N^2 \times precision) + recall}$$

Judged by F1, Metal surpasses all traditional callers (Fig. 5). With increasing priority given to recall, the impact of Scotch's superior recall becomes clear.

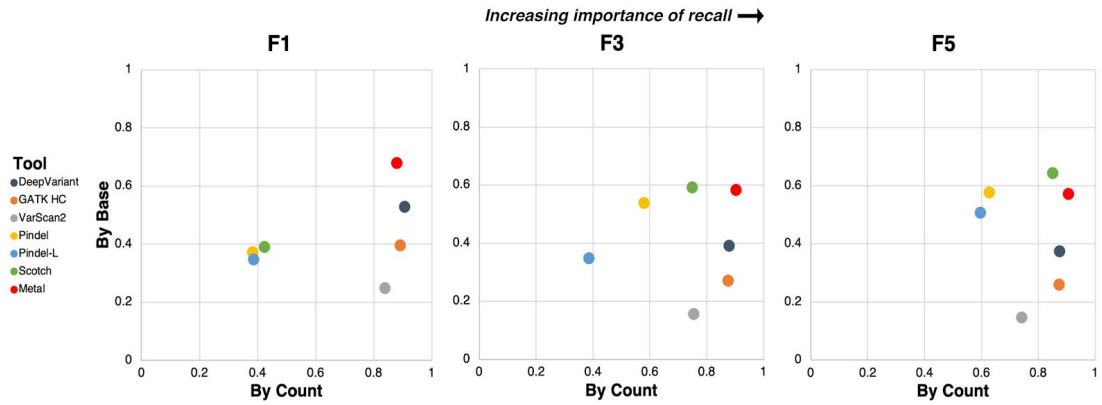

Fig. 5: F scores with recall by count and recall by base in Syndip.
We plot the F scores of the tools examined. In an F(N) score, recall is considered N times more important than precision. Recent callers such as DeepVariant and GATK HaplotypeCaller provide higher precision but are more likely to miss large variants. As the weight given to recall grows, Scotch and Metal surpass other callers.

Sanger sequencing validates variants not in consensus truth set, altering precision estimates
While more successful than other tools in identifying larger variants, Pindel-L, and, to a lesser extent, Scotch, exhibit low precision on the consensus truth set. On Syndip, the precision of DeepVariant, GATK HaplotypeCaller, and VarScan2 lies between 89% and 97%. On deletions, Scotch's precision is lower (72%) but higher than Pindel-L (27%) while on insertions, both Scotch (14%) and Pindel-L (3%) decline significantly.

We carried out Sanger sequencing to determine what proportion of variants identified as false positives by Scotch were, in fact, real variants absent from the consensus truth set. We selected for sequencing 100 indel breakpoint calls made by Scotch in NA12878 and flagged as false positives in GA4GH Benchmarking, with two constraints. First, half of the selected calls were insertion breakpoints, and half were deletion breakpoints (which, in turn, were half deletion-start and half deletion-end breakpoints). Second, 20 of the 100 calls were selected to have potential correlates in Syndip: indel calls within 3 bp, of any type, in the Syndip truth set. This constraint was introduced to determine whether Syndip had identified any additional common indels not detected by NA12878.

Analyzing resultant chromatograms with Poly Peak Parser[37], an online alignment-based tool that identifies indels, produced a validation rate of 35%. (The full results are available in Supplementary Table 21.) For 26 of the original 100 calls, surrounding GC content had been too high for effective primer design or PCR amplification failed. In an additional 20 cases, sequencing quality was too low to make an accurate determination. We further excluded 2 calls that were flagged as false positives not because they were missing altogether from the NA12878 truth set, but because Scotch had mis-identified their type. 18 out of the remaining 52 calls were verified as genuine indel breakpoints. 14 of these validated variants are homopolymer deletions.

If we consider these samples to be representative of Scotch's false positives, this implies that approximately 4321 of Scotch's reportedly false positive calls in fact refer to real variants missing from chromosome 22 of the NA12878 GiaB truth set. This result improves estimates of Scotch's precision. Based solely on the gold standard, we considered Scotch to have made 16,241 false positive calls across all indel breakpoints, out of 28,488 calls total. Reclassifying a

portion of former "false positives" variant calls as true positives, in proportion to validation rates from Sanger sequencing, improves precision from 43% to 58%.

These truth set validations also imply inflated values of recall for current callers. In line with Scotch's estimated 12247 true positive calls matching 9788 indel breakpoints, we estimate that, proportionately, 4321 of the Sanger-validated calls refer to 3453 new indel breakpoints. An "ideal" indel caller with recall and precision of 100% on the consensus gold standard, would not have identified any of these new true positives and its recall, upon consideration of these new variants, would decline to 74%. More information on these calculations is available in the Supplementary Note.

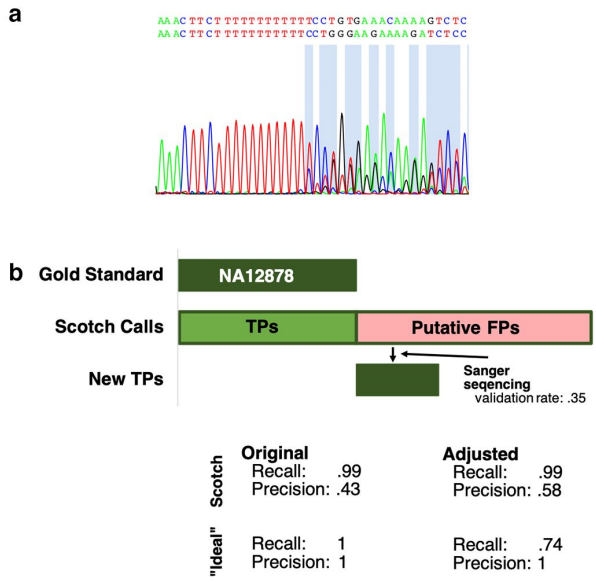

Fig. 6: Updated benchmarking following Sanger sequencing validation.
a, A sample chromatogram from Sanger sequencing of a call made by Scotch in NA12878 that was absent from the truth set. The mismatched peaks on the right side of the image indicate a heterozygous deletion.
b, A visual representation of the approximate relative magnitudes of the original truth set, the set of alleged false positives calls by Scotch, and the genuine variants within the alleged false positive calls that we expect from our Sanger sequencing. Below, metrics indicate updated estimates of the performance of Scotch and an "ideal" indel caller, which had a precision of 100% on the original truth set.

Scotch identifies clinically relevant variants missed by other tools

Our approach to improving indel calling was motivated by clinical application. While exome and genome sequencing are effective in diagnosing rare genetic disorders, estimates of diagnosis rate using this technology fall between 30% and 40%, leaving many patients undiagnosed. With Scotch, we sought to develop a method that would find these presumed genetic causes identifying as many true positives as possible, while minimizing the risk of missing the variant of interest.

We applied Scotch to the genomes of several patients presenting to Stanford's Undiagnosed Diseases Network (UDN). The UDN is a national consortium of medical centers taking on patients whose atypical constellations of symptoms have evaded diagnosis. We chose a

representative sample of undiagnosed patients at the Stanford center and applied the Scotch algorithm. In 21 of 26 cases, plausible candidates for the diagnosis were found. Such candidates would require further work to firmly establish disease causality but their presence in a large majority of undiagnosed cases is encouraging, especially considering that most variants detected in this way could be assumed to contribute to at least hemizygous loss of function. One example case illustrates the power of the new approach well.

An adult woman presented with distal asymmetric myopathy including scapular winging, mild facial weakness, decreased forced expiratory volume, and muscle biopsy notable for rimmed vacuoles and myofibrillar disorganization. In addition to a myopathy gene panel that was negative, whole-exome sequencing was performed for the patient, without a diagnosis. With whole genome sequencing data, Scotch made 4.5m indel breakpoint calls. (This is more than VarScan2, GATK HaplotypeCaller, and DeepVariant (1.1 - 1.9m), but fewer than Pindel (5.2m) and Pindel-L (12.7m).) Of Scotch's calls, 4,365 were deletion breakpoints within 100 bp of exons of ClinVar- and OMIM- annotated genes. 460 of these were seen in no unrelated samples, and a phenotype-based prioritization tool[38] ranked breakpoints of a 498-bp exon-skipping stop-loss deletion in *HNRNPA1* in rank 50. This deletion was not reported by DeepVariant, GATK HaplotypeCaller, or VarScan2; it was identified by Pindel, Pindel-L, and Metal. The implicated gene is a member of the hnRNP family, which has important roles in nucleic acid metabolism; mutations in *HNRNPA1* have been previously implicated in neuromuscular disease in patients with features which substantially overlap our case's phenotype.

Discussion

We present an approach to optimizing calling of insertions and deletions in the human genome. It is designed to optimize recall: evaluated on Syndip, across all indel breakpoints Scotch performs with higher sensitivity than any other tool. Scotch reports variants previously only accessible to long-read sequencing. Evaluated on simulated data, Scotch retains recall close to 1 on variants across the size spectrum. While high recall comes at the expense of lower precision, Sanger sequencing demonstrates that this is, to an extent, an artifact of a consensus dataset that omits some true positive indels. Taking these into account, the meta-calling approach incorporating Scotch surpasses all other tools and is likely to improve the diagnostic rate of clinical genome calling.

A significant advantage to benchmarking new algorithms is the recent publication and sharing of "reference" genomes derived from long read sequencing. These build on consensus datasets produced by the Genome in a Bottle consortium (which continues to expand its own reference collection in this direction). Notably, basic comparison of these "gold standard" callsets (NA12878 and Syndip) reveals major differences in the number and size distribution of variants too large to be explained by the diversity of individual human genomes. Syndip's use of long-read sequencing and multiple orthogonal variant callers provides for a greater number of variants that span a wider range of sizes, thus offering more comprehensive benchmarking opportunities.

Scotch's base-by-base procedure is less dependent on indel size than more coarse-grained approaches. For a caller that identifies variants by reconstructing regions of the genome through local assembly, the difference between a 40 bp and a 400 bp indel is significant. But while the complete presentation in sequencing data of these variants may differ, their breakpoints are

described by similar combinations of soft-clipped reads and changes in sequencing depth and quality. This relative conformity is the basis of Scotch's ability to detect indels of drastically different sizes. While trained primarily on NA12878 with a truth set produced by the Genome in a Bottle consortium, we added many large simulated indels to Scotch's training data to increase its sensitivity to large variants. Though trained only on indels of up to 500 bp, Scotch identifies variants in Syndip of up to several thousand base pairs in length.

We developed a meta caller (Metal) that delivers superior performance overall by integrating five variant callers. Collating the variants reported by these callers in its "smart intersection," Metal maximizes the number of true positive calls retained while filtering out erroneous calls resulting from sequencing errors. The high number of callers and their high initial sensitivity—as well as the loose comparison requirements—produces a master callset with high recall, including capture of many large variants, and high precision. Across all indel breakpoints in Syndip, Metal registers higher recall than any caller other than Scotch. At the same time, Metal greatly improves upon Scotch in terms of precision.

Sanger sequencing of variants called by Scotch missing from the NA12878 truth set reveals that some "false positives" are *bona fide* indels. Most of these are variants in homopolymer runs. While the reliability of Sanger sequencing itself declines in such regions, the prevalence of variants increases. We view improving sensitivity to these and other broader categories of variants as an imperative. We note that the incompleteness of benchmark datasets, in addition to presenting a challenge to machine-learning based approaches that may learn to miss the same indels, warps metrics derived from benchmarking. Considering the new true positives predicted by Sanger sequencing improves estimates of Scotch's precision, and lowers estimates of other callers' recall.

While gold standard datasets provide critical insight into the performance of variant callers, their potential for incompleteness means they should not be relied on exclusively. This is especially true in light of efforts to expand the capabilities of variant callers into broader categories of genetic variants—and those that lie in more challenging genomic regions—where current gold standard datasets are particularly likely to be incomplete. Over-reliance on benchmarking metrics may hinder the development of new tools by incorrectly penalizing improved callers with low precision, and rewarding those that maintain the "status quo" of identifying indels that are already confidently detected. In addition to evaluating callers against benchmark dataset, we encourage evaluation by Sanger sequencing of a sample of calls made outside the truth set for a more full picture of indel callers' capabilities. While attention to a variety of metrics is important, we urge greater focus on recall and improvement in discovery of a wider range of variants, relative to precision.

In summary, we present Scotch and Metal, tools capable of identifying new true positive insertion and deletion variants, expanding the range of variants that can be detected from next-generation sequencing data. We hope these tools, aided by insights from benchmark datasets, can continue to advance understanding of human disease and genetic diversity.

Methods

Input

Scotch accepts a Binary Alignment Mapping (BAM)[21] file containing next-generation whole-genome sequencing data. Scotch also accepts a FASTA file providing the corresponding reference genome. Scotch divides the input by chromosome for parallel processing.

Features

Scotch's model evaluates each position with respect to 39 features. These include "primary metrics," quantities which are extracted directly from sequencing data; "delta features" which track the differences in primary features between neighboring positions; and "genomic features," which describe the content of the reference genome at a given locus. Information on feature importance is available in the Supplementary Note (Supplementary Fig. 1, Supplementary Table 25).

Primary features

These 11 features are calculated directly from the sequencing data. Three describe coverage—including the number of reads, reads with no soft-clipping, and reads with a base quality of 13 or higher. Each of these are normalized across the sample for comparability with samples from various sequencing runs. Two more features describe the quality of the sequencing—the mean base quality and the mean mapping quality across all reads. Four more are calculated from the CIGAR string that details each read's alignment to the reference—recording the proportion of bases at that position across all reads that are marked as inserted, deleted, soft-clipped, and that at are at the boundary of soft-clipping (i.e., the base is soft-clipped but at least one neighboring base is not). Two more features describe the soft-clipping of the reads, if present: one gives the mean base quality of soft-clipped bases, another gives the *consistency score* of the soft-clipping.

A position's consistency score is a metric we derived that gives the ratio of the number of reads supporting the most common soft-clipped base (i.e., A, T, C, or G), to the number of all soft-clipped reads. Soft-clipping provides important signal of an indel to our model; this score helps a model distinguish indel-related soft-clipping (where all soft-clipped reads should support the same nucleotide) from that caused by low sequencing quality (where different nucleotides will be present).

Delta features

20 additional features give the change in each of the primary features listed above— except the soft-clipping consistency score—from a given locus to both of its neighbors.

Genomic features

Eight features, lastly, are derived from the reference genome, providing Scotch with insight into regions where sequencing errors are more common. Four of these features are binary: they indicate whether a genomic position is located in high-confidence regions, "superdup" regions, repetitive regions, and low-complexity regions. The remaining four describe GC-content (in windows of 50 and 1000 bp), mappability, and uniqueness.

Prediction and Output

These features are combined in a human-readable TSV that can serve as the input to any number of machine-learning setups. We trained several random forest models to identify the signals of

indels in this data. The primary output of Scotch is a VCF file that lists all breakpoints discovered, their confidence, and their type.

Training
We evaluated Scotch's performance when provided with labelled training data from five different sources: simulated variants, Syndip, CHM1, NA12878, and NA12878 with simulated variants. We found optimal results with the last source of training data. We also performed a hyperparameter optimization over random forest hyperparameters including the number of trees (ntree) and the number of predictors that can be considered at each node (mtry), though we found these to be largely insignificant.

Implementation
Scotch is implemented in Bash, Python, and R, and relies on the following packages: samtools, pysam, randomForest, and the GATK[20,30,39]. The codebase is publicly available on GitHub at https://github.com/AshleyLab/scotch, and the genomic features used by the machine learning model are available at https://github.com/AshleyLab/scotch-data, precomputed across the genome for convenience.

Acknowledgements
This work was supported by an award from the Stanford Center for Computational, Evolutionary, and Human Genomics. This project was also supported by contributions from the Jyotsna Sulebele Biomedical Data Science Fund. Research reported in this manuscript was also supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Number(s) U01 HG007708, U01 HG010218, U01 HG007530, and U01 HG007943. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by FDA Contract FDABAA-15-00121. This publication was also made possible by Grant number U01FD004979 from the FDA, which supports the UCSF-Stanford Center of Excellence in Regulatory Sciences and Innovation. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or FDA. R.L.G. was supported by a National Science Foundation Graduate Research Fellowship.

Author information
These authors contributed equally: Charles Curnin, Rachel L. Goldfeder.

Affiliations
Division of Cardiovascular Medicine, Stanford School of Medicine, Stanford, CA, USA
Charles Curnin, Shruti Marwaha, Daryl Waggott, Matthew T. Wheeler & Euan A. Ashley

Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
Rachel L. Goldfeder

Stanford Center for Undiagnosed Diseases, Stanford, CA, USA
Charles Curnin, Shruti Marwaha, Daryl Waggott, Matthew T. Wheeler & Euan A. Ashley

HudsonAlpha Institute for Biotechnology, 601 Genome Way, 35806, Huntsville, AL, USA
Elizabeth A. Worthey
Department of Genetics, Stanford School of Medicine, Stanford, CA, USA
Euan A. Ashley

Consortia
Undiagnosed Diseases Network
Maria T. Acosta, David R. Adams, Pankaj Agrawal, Mercedes E. Alejandro, Patrick Allard, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Guney Bademci, Eva Baker, Ashok Balasubramanyam, Dustin Baldridge, Deborah Barbouth, Gabriel F. Batzli, Alan H. Beggs, Hugo J. Bellen, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, David P. Bick, Camille L. Birch, Stephanie Bivona, Carsten Bonnenmann, Devon Bonner, Braden E. Boone, Bret L. Bostwick, Lauren C. Briere, Elly Brokamp, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Olveen Carrasquillo, Ta Chen Peter Chang, Hsiao-Tuan Chao, Gary D. Clark, Terra R. Coakley, Laurel A. Cobban, Joy D. Cogan, F. Sessions Cole, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Precilla D'Souza, Surendra Dasari, Mariska Davids, Jean M. Davidson, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Naghmeh Dorrani, Daniel C. Dorset, Emilie D. Douine, David D. Draper, Annika M. Dries, Laura Duncan, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Gregory M. Enns, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Irman Forghani, Noah D. Friedman, William A. Gahl, Rena A. Godfrey, Alica M. Goldman, David B. Goldstein, Jean-Philippe F. Gourdine, Alana Grajewski, Catherine A. Groden, Andrea L. Gropman, Melissa Haendel, Rizwan Hamid, Neil A. Hanchard, Nichole Hayes, Frances High, Ingrid A. Holm, Jason Hom, Alden Huang, Yong Huang, Rosario Isasi, Fariha Jamal, Yong-hui Jiang, Jean M. Johnston, Angela L. Jones, Lefkothea Karaviti, Emily G. Kelley, Dana Kiley, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Deborah Krakow, Donna M. Krasnewich, Susan Korrick, Mary Koziura, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, Byron Lam, Brendan C. Lanpher, Ian R. Lanza, C. Christopher Lau, Jozef Lazar, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Roy Levitt, Shawn E. Levy, Richard A. Lewis, Sharyn A. Lincoln, Pengfei Liu, Xue Zhong Liu, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Thomas C. Markello, Ronit Marom, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Thomas May, Jacob McCauley, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Jason D. Merker, Thomas O. Metz, Matthew Might, Eva Morava-Kozicz, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Avi Nath, Stan F. Nelson, J. Scott Newberry, John H. Newman, Sarah K. Nicholas, Donna Novacic, Devin Oglesbee, James P. Orengo, Stephen Pak, J. Carl Pallais, Christina GS. Palmer, Jeanette C. Papp, Neil H. Parker, John A. Phillips III, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Robb K. Rowley, Ralph Sacco, Jacinda B. Sampson, Susan L. Samson, Mario Saporta, Judy Schaechter, Timothy Schedl, Kelly Schoch, Daryl A. Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Kathleen Shields, Jimann Shin, Rebecca Signer, Catherine H. Sillari, Edwin K. Silverman, Janet S. Sinsheimer, Kathy Sisco, Kevin S. Smith, Lilianna Solnica-Krezel, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A.

Sullivan, David A. Sweetser, Cecelia P. Tamburro, Queenie K.-G. Tan, Mustafa Tekin, Fred Telischi, Willa Thorson, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Tiina K. Urv, Tiphanie P. Vogel, Daryl M. Waggott, Colleen E. Wahl, Nicole M. Walley, Chris A. Walsh, Melissa Walker, Jennifer Wambach, Jijun Wan, Lee-kai Wang, Michael F. Wangler, Patricia A. Ward, Katrina M. Waters, Bobbie-Jo M. Webb-Robertson, Daniel Wegner, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Jeremy D. Woods, Elizabeth A. Worthey, Shinya Yamamoto, John Yang, Amanda J. Yoon, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, Stephan Zuchner.

Contributions
C.C. implemented the pipeline, analyzed its performance, and drafted the manuscript. R.L.G. developed the approach including the initial pipeline and feature selection. S.M., D.W., M.T.W., and E.A.A., provided statistical and computational support and supervision. R.L.G., C.C., D.W., M.T.W., and E.A.A. designed the experiments and approach. R.L.G., M.T.W., and E.A.A. secured funding. All authors provided critical review and input into the final content of the manuscript.

Competing interests
M.T.W. and E.A.A. report holding stock in Personalis, Inc. E.A.A is a founder of Personalis, Inc. and Deepcell Inc. EAA is an advisor to Sequence Bio and Genome Medical.

Corresponding author
Correspondence to Euan Ashley euan@stanford.edu

References

1. Wu, M., Chen, T. & Jiang, R. Leveraging multiple genomic data to prioritize disease-causing indels from exome sequencing data. *Sci. Rep.* **7**, (2017).

2. Manta, F. S. N. *et al.* Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. *Ann. Hum. Biol.* **40**, 94–98 (2013).

3. Ng, P. C. *et al.* Genetic Variation in an Individual Human Exome. *PLoS Genet.* **4**, e1000160 (2008).

4. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and

next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).

5. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5**, 17875 (2015).
6. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880–1887 (2014).
7. Yang, Y. *et al.* Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA* **312**, 1870 (2014).
8. Merker, J. D. *et al.* Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**, 159–163 (2017).
9. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 1 (2019).
10. Hasan, M. S., Wu, X. & Zhang, L. Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics* **9**, 20 (2015).
11. Dewey, F. E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035–1045 (2014).
12. Grimm, D., Hagmann, J., Koenig, D., Weigel, D. & Borgwardt, K. Accurate indel prediction using paired-end short reads. *BMC Genomics* **14**, 132 (2013).
13. Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr. Protoc. Bioinformatics* **45**, 15.6.1–11 (2014).
14. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).

15. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

16. Narzisi, G. *et al.* Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**, 1033–1036 (2014).

17. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

18. Shigemizu, D. *et al.* IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Sci. Rep.* **8**, 5608 (2018).

19. Yang, R., Nelson, A. C., Henzler, C., Thyagarajan, B. & Silverstein, K. A. T. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome Med.* **7**, 127 (2015).

20. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

21. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

22. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

23. Yang, J. *et al.* InDel marker detection by integration of multiple softwares using machine learning techniques. *BMC Bioinformatics* (2016). doi:10.1186/s12859-016-1312-2

24. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark

somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).

25. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
26. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
27. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**, 3694–3696 (2015).
28. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).
29. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
30. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
31. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
32. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
33. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
34. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4235

35. Krusche, P. *et al.* Best Practices for Benchmarking Germline Small Variant Calls in Human Genomes. (2018). doi:10.1101/270157

36. Guo, Y., Ding, X., Shen, Y., Lyon, G. J. & Wang, K. SeqMule: automated pipeline for analysis of human exome/genome sequencing data. *Sci. Rep.* **5**, 14283 (2015).

37. Hill, J. T. *et al.* Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev. Dyn.* **243**, 1632–1636 (2014).

38. Birgmeier, J. *et al.* AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. *bioRxiv* 171322 (2017). doi:10.1101/171322

39. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. in *Current Protocols in Bioinformatics* 11.10.1–11.10.33 (2013).