Opinion

CellPress

# Accounting for uncertainty in DNA sequencing data

Jason A. O'Rawe[1,2,3], Scott Ferson[2,3], and Gholson J. Lyon[1,2,4]

[1] Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA
[2] Stony Brook University, Stony Brook, NY, USA
[3] Applied Biomathematics, Setauket, NY, USA
[4] Utah Foundation for Biomedical Research, Salt Lake City, UT, USA

**Science is defined in part by an honest exposition of the uncertainties that arise in measurements and propagate through calculations and inferences, so that the reliabilities of its conclusions are made apparent. The recent rapid development of high-throughput DNA sequencing technologies has dramatically increased the number of measurements made at the biochemical and molecular level. These data come from many different DNA-sequencing technologies, each with their own platform-specific errors and biases, which vary widely. Several statistical studies have tried to measure error rates for basic determinations, but there are no general schemes to project these uncertainties so as to assess the surety of the conclusions drawn about genetic, epigenetic, and more general biological questions. We review here the state of uncertainty quantification in DNA sequencing applications, describe sources of error, and propose methods that can be used for accounting and propagating these errors and their uncertainties through subsequent calculations.**

## All uncertainties are not created equal

Personalized and genomics-guided medical care promise to revolutionize the way that we treat and prevent human disease by relying more heavily on accurate and rich characterizations of individuals, rather than on population-scale phenomenon. Thus, it is becoming increasingly important and relevant for analysis methods to guarantee rigorous accounts of individuals, whose medical treatment will be shaped by these results. At the forefront of personalized medicine is DNA sequencing, and relatively standardized methods for analyzing these data have been developed [1–6]. As sequencing becomes more routine in the clinic, it is important to consider the accuracy of these data and the validity of the conclusions based on them.

DNA-sequencing data contain two major types of quantitative uncertainty, which we refer to here as aleatory and epistemic. Aleatory uncertainty refers to the variability that is inherent to most biological systems, such as stochastic fluctuations in a quantity through time or variation across space. This is considered to be a form of uncertainty because the value of the quantity can change each time a measurement is taken, and we cannot predict precisely what the next value will be [7]. In the context of DNA-sequencing studies, variability can arise from sequencing DNA that has been extracted from tissues with differing genotypes, sequencing large populations for the purpose of determining population allele frequencies, and from varying RNA expression levels across space or time, just to name a few sources. By contrast, epistemic uncertainty refers to incomplete knowledge about a quantity, which can arise from imperfect measurement, limited sampling effort, or ignorance about the underlying processes that influence a quantity [7]. Again, in the context of DNA sequencing, this type of uncertainty can result from poor base detection, sparse sequence data, or from a fundamental lack of understanding about how sources of error arise in these data and how they are related.

These two forms of uncertainty have important practical differences. For example, epistemic uncertainty can in principle be reduced by empirical effort and, although aleatory uncertainty (variability) can sometimes be better characterized through repeated experimentation, it cannot generally be reduced by empirical effort. The differences between these two forms of uncertainty are often significant in practical settings.

Substantial progress has been made toward quantifying and propagating uncertainty through calculations and inferences made on human DNA-sequencing data, and this progress has already yielded more accurate characterizations of population-scale phenomenon [8–13]. However, these methods are limited in that they are not easily extended for use in the many different, and often times piecemeal, analyses that are currently necessary for implementing personalized genomics. Furthermore, the treatment of uncertainty is currently limited in scope, due to the difficulties inherent in modeling different sources and types of uncertainty that often arise in personalized genomics-based analyses. A general framework for propagating uncertainty through calculations is needed, so that computations can be made with as much rigor as is possible, given the available technologies.

Here, we describe sources of error that arise in high-throughput sequencing data, describe components of these errors that are the consequence of manifestations of different types of uncertainty, advocate for the development of

*Corresponding author:* O'Rawe, J.A. (jazon33y@gmail.com).
*Keywords:* DNA sequencing; uncertainty; sequence errors; uncertainty accounting.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2014.12.002

new DNA-sequencing analysis methods for computing with uncertain data in the context of personalized genomics applications, and describe prospective methods that allow for incorporation of uncertainties into their computational frameworks so that genomic inferences can more accurately represent the true state of knowledge.

## Sources of uncertainty in DNA sequencing

DNA consists of categorically defined units, nucleotide bases. Nucleotides, or sequences of nucleotides, consist primarily of adenine (A), guanine (G), thymine (T), or cytosine (C) bases. In practice, the reliability of DNA sequence detection varies from base to base and is usually influenced by the specific sequencing technology used. High-throughput and short-read sequencing technologies generally quantify detection reliability using probability values that characterize the chance that a base was correctly detected [14,15]. This value depends in part on the chemistry used in the sequencing, the particular equipment used to detect DNA, and sequence composition, among other things. Relatively short sequence reads (approximately 150 consecutive bases) are generated through iterative and consecutive base detection, which are then aligned to a reference sequence. The alignment procedure also generally produces probability values that characterize the chance of having correctly aligned a sequence to its respective genomic location. These two values taken together give the analyst information about the chance that a base is correct.

Errors in high-throughput sequencing data (base calling errors, variant calling errors, etc.) result from complicated and sometimes unforeseen technical and data-processing factors. Several empirical studies have identified various different important and quantifiable sources of error. In a recent perspective piece [16], the authors reviewed these different sources, which include upstream steps during sample preparation and sequence library preparation, as well as from the sequencing, imaging, data-processing, and bioinformatics steps [17]. More specifically, errors originating from sample preparation are sometimes due to a combination of human errors in sample handling (which can include sample swaps or DNA and/or RNA degradation), sample contamination, and low quantities of input DNA. During the preparation of sequence libraries, human errors can result in cross-contamination of DNA samples across different library preparations, and errors can occur when PCR amplification incorporates an incorrect base during early synthesis cycles [16]. Primer-mediated sequence amplification biases, the synthesis of chimeric reads, barcode or adapter errors, and machine failures are among the other sources of error that originate during sequence library preparation. During base imaging and sequencing, user errors combined with the incorporation of additional bases during single sequence cycles, DNA damage, overlapping signals, strand biases, sequence complexity [18], and machine failures can contribute to sequence error. Moreover, errors in bioinformatics steps resulting from poor sequence alignment in regions where mapping is difficult [19] also contribute to sequence error.

Analysts use a variety of algorithmic and statistical approaches for mitigating errors and for quantifying uncertainties about DNA sequence related estimates. Contemporary tools leverage efficient sequence alignment-based frameworks for detecting similarities and differences between sample and reference sequences [20,21]. Initial analysis steps entail excluding putative sequence error or low-quality sequences using data-quality thresholds (i.e., if a sequence is of $x$ or less quality, then exclude it from the analysis). Variations of the Smith–Waterman [22] algorithm are then used to match and align similar sequences. Once the sample sequence has been aligned to the reference sequence, various statistical approaches are used to identify the most likely genotype, including Bayesian inference [21,23], frequentist hypothesis testing [24], and others [25]. More recent algorithmic enhancements use local sequence assembly to mitigate errors caused by aberrant alignments and to detect complicated sequence differences between the sample and the reference sequence [26–28]. ‘Error correction’ and hybrid-sequencing approaches generate high-quality sequence data by correcting error-prone long read technologies with high-fidelity short read sequencing technologies [29,30]. These error correction techniques enable reference-free assemblies of larger genomes, reduce sequence alignment artifacts, and allow for the sequencing of genomes with no known reference sequence.

There are two major deficiencies in current analysis approaches with respect to the appreciation of uncertainty and errors in DNA sequencing data. The first deficiency is that epistemic uncertainties are currently not well quantified. The second is that uncertainty, even if quantified, is often not properly incorporated into subsequent analyses and calculations. These limitations are compounded by the fact that there are no software implementations that allow for uncertainty quantifications to be carried through and used in routine downstream calculations and analyses. Moreover, software libraries for computing with epistemically uncertain data are almost nonexistent and are not widely available to most practicing bioinformaticians.

We view the accurate representation and incorporation of uncertainties into DNA-sequencing analyses as a necessary piece of a more general computational solution that generates full, honest, and robust determinations of the reliabilities of inferences stemming from these data. It is important to recognize that heuristic filtering approaches essentially discard imprecise or poorly collected and/or understood data, but these data should and can be included in the analysis.

## Quantifying uncertainty

Among the various sources of error and uncertainty discussed above, those that are known to arise through random processes or are a result of system-level variability can be modeled using established statistical approaches. They are often aleatory in nature because more sequencing data will not reduce the resulting allelic variability; it would simply result in a more precise characterization of it.

When errors are not known to be random and could instead be systematic in nature, then established statistical approaches may not always conveniently apply. Systematic errors are due, in some cases, to inaccuracies in instrument calibration or to a bias in the way that the instrument takes particular measurements [31]. As just one example in the context of DNA sequencing in personal genomics applications, if either the reverse or forward DNA strand is sequenced more than one would expect through a random sampling of each, it is considered evidence of potential sequence bias. This type of bias leads to uncertainty about the underlying sequence composition that is epistemic in nature; the analyst simply does not have representative samples of both DNA strands. In these situations, the analyst has less or sometimes even no information about what the other strand looks like in terms of its allelic composition. This creates epistemic uncertainty about the true genetic sequence and, therefore, one cannot reasonably make any distributional assignments in lieu of the missing data.

In situations where distributional assignments are not well justified, epistemic uncertainties can often be conveniently modeled using intervals [32]. Intervals are a reflection of our inability to assign any distributional information to the various possible states of some variable of interest. The state-of-the-art in DNA sequence analysis simply discards or ignores data that show evidence of systematic error. However, this is unnecessarily strict and throws away potentially useful information. In the case of sequence strand bias, one conservative strategy might be to construct an interval (in this case, a set) that includes all possible bases. Although such an interval is a quantitative admission of ignorance, it can accurately represent the epistemic uncertainty that results from one strand being sequenced less often than expected. This uncertainty concerns one of the two DNA strands, the other of which may be well characterized. Instead of discarding unreliable data, characterizing its uncertainty and carrying these uncertain data through analyses, although not always resulting in something easily interpreted, can in many cases be useful. It remains an open question as to how one should model systematic uncertainties in sequence data. Schemes are needed for understanding and better modeling systematic uncertainties so that they can be accurately quantified and propagated through calculations.

In the subsections below, we suggest two, of many possible, aspects of DNA-sequencing analysis where improvements in quantifying and propagating uncertainties can be achieved using existing computational tools. We provide two hypothetical situations, in Boxes 1 and 2, which exemplify how analysts might perform these computations in situations of pervasive uncertainty. These methods and examples should not be taken as comprehensive solutions to the problem at hand, but rather propositions. The challenge of accurately quantifying both epistemic and aleatory uncertainties in DNA sequence-related data sets and then subsequently propagating them through analyses is not a solved problem. We hope to stimulate some discussion so that comprehensive solutions might, in the future, be found.

### *Uncertainty about dependencies*

Uncertainties about dependency relations are an important consideration when combining sources of error for making inferences based on high-throughput DNA sequencing data, because errors can result from varied or shared processes. As an example, multisample variant callers use information spanning many different samples to generate calls. This means that the variant detection between individual samples is not a completely independent process, and so performing logical operations on these variants using information about their probability of being correct cannot be done assuming their independence (Box 1). When knowledge about the dependency relation between two variables is

**Box 1. Probability of shared alleles under conditions of uncertainty**

Genomic variants can act as predictors of disease, disease progression, and outcome. In cases of inherited genetic disease, even seemingly trivial uncertainty calculations may make an important difference.

For instance, imagine a case where a pathogenic variant has been detected in a mother with a 0.9 probability of being correct, but was detected in the son with a 0.5 probability of being correct (Figure I). Typically, analysts filter out low-quality variant calls, which, in this case, would result in an analysis that says the mother and son do not share this variant. Instead of filtering the lower-quality variant, one could instead calculate the chance that this variant is present in both the child and the mother. If we assume that the variant detections for the mother and son are independent, then the probability that they both have the variant is $0.9 \times 0.5 = 0.45$.

In many practical sequencing applications, variants are detected using information spanning multiple related or unrelated cohorts, so genotype inferences are no longer made in an entirely independent manner. Performing the same logical operations, but instead assuming nothing about the dependence relation between the two estimates, results in the simple conjunction 0.5 & 0.9 degenerating into an interval answer, in this case [0.4, 0.5].

It is important to note that for pedagogical purposes, we assumed in this example that the variant quality scores (which are translated into probability values) were computed with absolute precision, and that they accurately represent uncertainty about the call. In practice, this is rarely if ever the case, despite the fact that they are almost always reported as such. Indeed, Phred quality scores that are less than 30 are considered unreliable, which translates into a 0.999 lower bound on variant calling accuracies. These quality scores have been shown to underestimate the probability of variant calling errors made by various different variant-calling algorithms [17,27]. This is perhaps a more pervasive and important issue needing attention from the field, although some statistical and algorithmic approaches have been developed for generating more accurate quality scores [21,64,65].

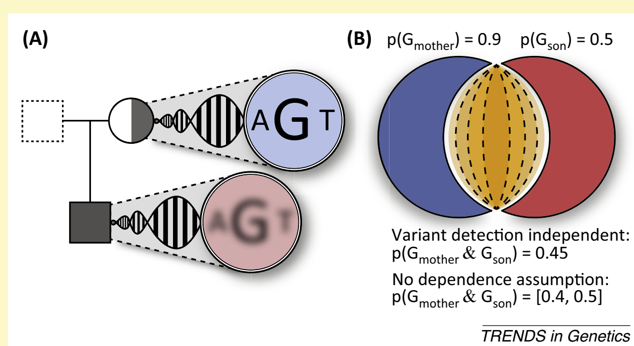


**Figure I.** Here, we demonstrate how uncertainty about dependencies results in requiring the necessary framework for computing with interval values, which are quantitative representations of epistemic uncertainty. Uncertainty about variant detection depicted as blurring **(A)** and encoded as probabilities that are logically conjoined **(B)** using calculations that assume independence or calculations that make no assumptions about their relation.

Trends in Genetics xxx xxxx, Vol. xxx, No. x

**Box 2. Statistical inference under conditions of pervasive uncertainty**

A common task in sequence analysis is to determine sequence composition in single- or multisample data sets. This task is often made difficult by sparse data sets, a lack of prior knowledge for use in statistical inferences, and the presence of imprecise or otherwise uncertain input data (due to, for example, noisy or poor raw sequence data).

Imagine a data set comprising sparse DNA sequences generated from an individual of a never-before sequenced population. The analyst is initially faced with two distinct problems inherent in the particularities of the experiment: sparse data and no prior information about expected allele frequencies. Suppose the available data come from eight sequence reads, and we observe {T, T, A, T, T, T, T, T} at a particular locus (Figure I), with each base characterized as being detected with some degree of uncertainty, due either to known systematic error or from combining estimates of error from multiple sources. The analyst is now faced with a third, and perhaps more difficult, analysis challenge: figuring out how to incorporate epistemic uncertainty about base detection in the quantification of allele frequencies.

For the sake of simplifying this example, let us say that the bases are recorded with accuracies represented by interval probabilities: {[0.9, 0.99], [0.8, 0.9], [0.4, 0.6], [0.8, 0.9], [0.98, 0.99], [0.7, 0.9], [0.4, 0.9], [0.8, 0.85]}, respectively. An analyst could then use a simple model that computes allele frequencies as probabilities from a multinomial distribution. With each base itself modeled as a Bernoulli process with interval probabilities, a confidence structure on allele frequencies can be obtained using the formula $I_x([k_A, k_A+1], [n-k_A+1, n-k_A])$, where $k_A$ and $n$ are the number of observed alleles of interest out of the total, respectively, and $I_x$ is the regularized incomplete beta function [66]. The resulting confidence structure does not assume or require prior knowledge about the expected allele frequencies and we can, from it, compute confidence intervals about the estimate. In particular, we can say with 95% confidence that the allele frequency of adenine at this locus is between 0.002 and 0.526. This broad range reflects the fact that there were only eight data samples, no prior allele frequency information, and uncertainty about each base call.

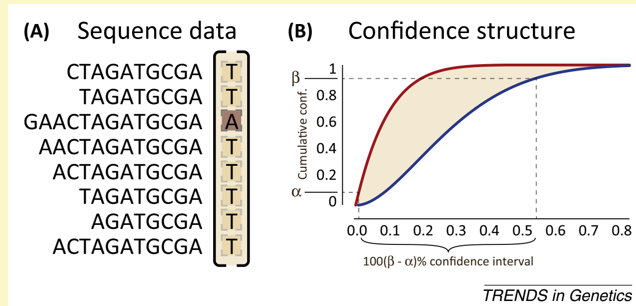


**Figure I**. Here, we show how variability in sequence composition can be characterized using a statistical inference routine that can compute over epistemically uncertain input data, which in this case refers to interval probability values characterizing base calling errors. Sequence data **(A)** are used to estimate the frequency of an allele **(B)**, in this case for adenine, at a particular locus using a multinomial model and confidence structures for statistical inference.

unavailable, their conjunction (e.g., the probability of the joint event A & B) can be computed using the formula $[\max(0, a+b-1), \min(a, b)]$, where $a$ represents th]e probability of some event, A, occurring and $b$ represents the probability of another event, B, occurring [32,33]. Similarly, the logical disjunction can be computed using the formula $[\max(a, b), \min(1, a+b)]$. These represent the tightest possible bounds on calculations when we do not know the dependence; they quantify our epistemic uncertainty about the underlying relation between A and B. It is in these cases that we are left to compute with intervals or even imprecise probability distributions (if the input operands are distributions), rather than simple point values and precise distributions. Interval analysis [34–38] is the simplest method for performing arithmetic and logical operations on interval data, and the results can be made as precise as possible given the input values. Probability bounds analysis [7,39–42] allows quantities with epistemic uncertainty represented by intervals to be combined with random distributions representing aleatory uncertainty in mathematical expressions. This method is part of the theory of imprecise probabilities [43]; it allows calculations when only bounds on the input distributions are known.

### *Model uncertainty*

Statistical models inform most DNA analysis algorithms, but this task is often made difficult by sparse data sets, a lack of prior knowledge for use in inferences, and the presence of imprecise or otherwise uncertain input data (due to, for example, noisy or poor raw sequence data). Methods that can make inferences in the presence of these complicating difficulties are needed.

Robust Bayesian [44–49] inference allows for the consideration and analysis of imprecise sample data or ignorance about the appropriate prior assumptions (both of which are the manifestations of epistemic uncertainty and can be modeled using intervals, bounding approaches, and robust Bayesian approaches). In a robust Bayesian analysis, results are considered robust if neither imprecision in the input data nor differences in the prior probability distribution have large effect on the output.

Similar to robust Bayesian approaches, confidence structures [50–52] characterize inferential uncertainty about statistical estimates from sparse or imprecise data, but the confidence structure approach does not require the use of prior knowledge. Confidence structures are similar to Bayesian posterior distributions because they estimate distribution parameters from sample data. They give us confidence intervals for all levels of confidence (Box 2) and, with them, analysts can guarantee statistical performance through their repeated use. Confidence structures are useful in practical applications because they can be used in arithmetic or logical calculations, and the results will still guarantee statistical performance; that is, they will still yield true confidence intervals.

## Concluding remarks

As DNA sequencing technologies migrate from science to commerce, incentives to publish full accounts of error rates diminish. Even if commercial entities have access to large-scale in-house sequence validation data, proprietary and business interests may prevent open publication of these critical data. Commercial entities now routinely provide fee-for-analysis services, but cannot, or otherwise choose not to, release specific information about data-processing procedures used in the analyses. Although practically useful, and perhaps even reliable for the problems at hand, this practice fundamentally violates basic conventions required to call the endeavor science. Furthermore, many of the most accurate and most widely used software tools for sequence alignment and variant detection are, in fact, closed source and their algorithms are described only in general terms without explicit, published definitions (e.g., Novoalign [53] or GATK HaplotypeCaller [54]). As such, they are not available even to scientists working on applied problems. In principle, software tools can be partially validated against synthetic data whose true nature is known and specifically described, but validation experiments performed on real, and often more complex, data are generally necessary for making post-hoc determinations of sequencing error rates. Indeed, methods that perform similarly with synthetic data may perform decidedly differently on real data [27]. Thus, costly empirical work is generally needed for better characterizing specific or more global sequencing error rates arising from the various sources of error.

Our view is that error estimates and related uncertainties should be incorporated into analyses as they arise, from uncertainties in initial measurements to uncertainties and errors from combinations and mathematical manipulations of data and inferences. This strategy allows for dynamic control over error rates and false findings, and it leverages the available data for responsive and real-time estimates of putative errors. We view uncertainty propagation and post-hoc error quantifications as complementary approaches, and many studies have found practical use in quantifying errors and accuracies of inferences in a post-hoc manner [27,55,56]. Here, we have focused on describing uncertainties arising in detecting DNA sequences and in differences between sample and reference sequences, but other techniques uncover information on higher-order biological phenomenon. One such example is Hi-C [57], which is a method that allows researchers to better understand the 3D organization of DNA in the cell [57,58], microbial community composition [59,60], and others [61]. Higher-order inferences should be made with algorithmic and statistical strategies that allow for the full appreciation of uncertainties in underlying measurements and determinations, because validating these inferences can sometimes be difficult due to the complexity of these systems and the rarity and precious nature of the underlying biological samples.

There is a need for rigorous uncertainty accounting across DNA detection and downstream sequencing-related analyses. As we progress through a scientific era in which nucleotide resolution studies are becoming the normal means of genetic dissection, science has found that even single nucleotide mutations can result in serious human disease [62,63]. Therefore, it is dangerous for an analysis to be wrong about any given variant call, particularly for biological samples that cannot be used for secondary studies or for orthogonal validations. Robust and full uncertainty accounting allows the analyst to better understand and predict when data and inferences are reliable and when they are not. This in turn informs data collection efforts, improves the reliability of published biological inferences based on error-prone DNA-sequencing technologies and improves the quantitative rigor associated with all analyses based on DNA sequencing.

## References


1 Anders, S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols* 8, 1765–1786

2 Trapnell, C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* 7, 562–578

3 Cooper, G.M. and Shendure, J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 12, 628–640

4 Nielsen, R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451

5 Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342

6 Hawkins, R.D. *et al.* (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486

7 Ferson, S. *et al.* (2002) *Constructing Probability Boxes and Dempster–Shafer Structures*, Sandia National Laboratories

8 Skotte, L. *et al.* (2012) Association testing for next-generation sequencing data using score statistics. *Genet. Epidemiol.* 36, 430–437

9 Vieira, F.G. *et al.* (2013) Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome Res.* 23, 1852–1861

10 Skotte, L. *et al.* (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195, 693–702

11 Fumagalli, M. *et al.* (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* 195, 979–992

12 Korneliussen, T.S. *et al.* (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* 14, 289

13 Kim, S. *et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12, 231

14 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194

15 Richterich, P. (1998) Estimation of errors in 'raw' DNA sequences: a validation study. *Genome Res.* 8, 251–259

16 Robasky, K. *et al.* (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62

17 O'Rawe, J. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 5, 28

18 Purcell, S.M. *et al.* (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190

19 Lee, H. and Schatz, M.C. (2012) Genomic dark matter: the reliability of short read mapping illustrated by the Genome Mappability Score. *Bioinformatics* 28, 2097–2105

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

21 DePristo, M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498

22 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197

23 Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv* 1207, 3907

24 Wei, Z. *et al.* (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next–generation sequencing data. *Nucleic Acids Res.* 39, e132

25 Zhao, Z. *et al.* (2013) An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Ann. Appl. Stat.* 7, 2229–2248

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

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

28 Li, S. *et al.* (2013) SOAPindel: Efficient identification of indels from short paired reads. *Genome Res.* 23, 195–200

29 Lee, H. *et al.* (2014) *Error correction and assembly complexity of single molecule sequencing reads*. Published online June 18, 2014. (http://dx.doi.org/10.1101/006395)

30 Koren, S. *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700
31 Meacham, F. *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12, 451
32 Ferson, S. *et al.* (2007) *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories
33 Fréchet, M. (1935) *Généralisations du Théorème Des Probabilités Totales*, Fundamenta Mathematica
34 Neumaier, A. (1990) *Interval Methods for Systems of Equations*, Cambridge University Press
35 Alefeld, G. and Herzberger, J. (1984) *Introduction to Interval Computation*, Academic Press
36 Moore, R.E. and Moore, R. (1979) *Methods and Applications of Interval Analysis*, SIAM
37 Moore, R.E. (1966) *Interval Analysis*, Prentice-Hall Englewood Cliffs
38 Dwyer, P.S. (1951) *Linear computations*, John Wiley, (New York)
39 Frank, M.J. *et al.* (1987) Best-possible bounds for the distribution of a sum: a problem of Kolmogorov. *Probab. Theory Rel. Fields* 74, 199–211
40 Williamson, R.C. and Downs, T. (1990) Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *Int. J. Approximate Reasoning* 4, 89–158
41 Ferson, S. (1995) Quality assurance for Monte Carlo risk assessment. In *Uncertainty Modeling and Analysis, 1995, and Annual Conference of the North American Fuzzy Information Processing Society. Proceedings of ISUMA-NAFIPS'95, Third International Symposium*, pp. 14, 19, 17–19. IEEF (http://dx.doi.org/10.1109/ISUMA.1995.527662)
42 Ferson, S. and Hajagos, J.G. (2004) Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliab. Eng. Syst. Saf.* 85, 135–152
43 Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall London
44 Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer
45 Berger, J.O. *et al.* (1994) An overview of robust Bayesian analysis. *Test* 3, 5–124
46 Insua, D.R. and Ruggeri, F. (2000) *Robust Bayesian Analysis*, Springer
47 Pericchi, L.R. (1998) *Sets of Prior Probabilities and Bayesian Robustness*, The Society for Imprecise Probability: Theories and Applications (http://www.sipta.org/documentation/robust/pericchi.pdf)
48 Pericchi, L.R. and Pérez, M.E. (1994) Posterior robustness with more than one sampling model. *J. Stat. Plann. Inference* 40, 279–294
49 Moreno, E. and Pericchi, L.R. (1993) Bayesian robustness for hierarchical ε-contamination models. *J. Stat. Plann. Inference* 37, 159–167
50 Ferson, S. *et al.* (2014) Computing with confidence: imprecise posteriors and predictive distributions. In *Vulnerability, Uncertainty, and Risks: Quantification, Mitigation, and Management* (Beer, M. *et al.*, eds), pp. 895–904 ASCE (http://dx.doi.org/10.1061/9780784413609.091)
51 Balch, M.S. (2012) Mathematical foundations for a theory of confidence structures. *Int. J. Approximate Reasoning* 53, 1003–1019
52 Ferson, S. *et al.* (2013) Computing with confidence. In *Proceedings of the Eighth International Symposium on Imprecise Probability: Theory and Applications*. Compiègne University, France
53 Novocraft (2014) *Novoalign*, Novocraft (https://www.broadinstitute.org/gatk/guide/article?id=4146)
54 Broad Institute (2014) *HaplotypeCaller*, Broad Institute (https://www.broadinstitute.org/gatk/guide/article?id=4146)
55 O'Rawe, J. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 5, 28
56 Fang, H. *et al.* (2014) Reducing INDEL calling errors in whole-genome and exome sequencing data. *Genome Med.* 6, 89
57 Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293
58 Dixon, J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380
59 Beitel, C.W. *et al.* (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* 2, e415
60 Burton, J.N. *et al.* (2014) Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability Maps. *G3* 4, 1339–1346
61 Sanyal, A. *et al.* (2012) The long-range interaction landscape of gene promoters. *Nature* 489, 109–113
62 Lyon, G.J. and Wang, K. (2012) Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Med.* 4, 58
63 Rope, A.F. *et al.* (2011) Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am. J. Hum. Genet.* 89, 28–43
64 Nielsen, R. *et al.* (2012) SNP Calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* 7, e37558
65 Van der Auwera, G.A. *et al.* (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons
66 Pearson, K. (1968) *Tables of the Incomplete Beta-Function*, Cambridge University Press