SOUND
Statistical Multi-Omics Understanding

Project Deliverable

Project acronym: **SOUND**	GA number: **633974**
Project title: **Statistical Multi-Omics Understanding of Patient Data**	
Funding Scheme: Collaborative Project (H2020-PHC-2014-2015/H2020-PHC-2014-two-stage) Health, novel medical developments	
Project start date: **01 September 2015**	Duration: **36 months**
Project's coordinator: Dr Wolfgang Huber (European Molecular Biology Laboratory, Heidelberg)	

D11.1 Technical report with best practice guidelines and software package

Due date of deliverable: Month 18 - 28.02.2017
Actual submission date: 23.02.2017

Organization name of lead contractor for this deliverable: University of Cambridge (UCAM)
Organization name of other involved partners: EMBL, ETH, TUM, IDMEC and USZ
Personnel involved: *Andy Lynch Schüler and Daniele Biasci*

Project co-funded by the European Commission within the H2020 Program (2015-2018)		
Dissemination Level		
PU	Public	x
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Deliverable description and summary

A report comprising part of deliverable 11.1 in the SOUND project, addressing as it does the practice, theory and implications of mutation calling in cancer sequencing projects - in particular we discuss characterizations of the performance of mutation calling such as false positive and false discover rates and their limitations, while proposing an interpretation of recall that avoids some of the limitations and offers extensions to experimental design (both primary sequencing and validation experiments). This report is accompanied by a prototype software package (https://github.com/db544/varmiss) implementing the proposed approach that, after development in accordance with other SOUND work packages, will be published as a full Bioconductor[1] package.

[1] http://www.bioconductor.org

Reporting Uncertainty in Somatic Mutation Calling

Prepared by Andy Lynch [UCAM] following discussions with project members.

Overview

Under the hypothesis that cancer is caused by mutations in the genome, it is natural that studies such as ICGC and TCGA wish to catalogue those mutations and look for ‘driver’ events. What is less clear is how well we need to perform this task for any individual sample (in the context of a population study) nor indeed how well we can assess that performance. For simplicity we will consider the Single Nucleotide Variant (SNV) as our mutation type in this report and consider the motivation that we want to identify genes that are recurrently somatically mutated in a population of patients with (presumable the same type of) cancer.

Other types of mutations (e.g. InDels, Structural Variants) exist, and indeed other motivations (e.g. in-depth understanding of a patient’s case, or wanting to find co-mutated genes) exist, but we will address these only in passing. There are also different motivations for assessing the performance of SNV calling: we may be assessing the mutation-calling-tool or be assessing the usefulness of our call set - and these again may require different interpretation.

For an overview of the challenges of SNV calling, see Alioto *et al.* [1].

Background to SNV and recurrent Gene calling:

In principle we should have an easy task. Given sequencing of normal tissue, and sequencing of the tumour, we simply need to identify those locations where different nucleotides are reported. There are a number of factors that play into this.

Issues 1: The samples as representations of their parent tissues.

The theme of the recent trend for molecular characterisation of tumours might be summarized as heterogeneity on virtually every level investigated. This has a direct impact on the ability to meaningfully interrogate a cancer via a single sample. Put simply, there is a tension between capturing the underlying heterogeneity by sampling a large region, but producing data that represent well an average of tumour behaviour but no individual cells, or a small number of cells (at the extreme, a single cell) that provide a truth for the tissue that they represent, but tell nothing of the rest of the cancer.

Drawing DNA from a large sample, and the inherent heterogeneity within that sample, will reduce the allele fraction of any mutations not present in all cells (which may be the majority). Such mutations may be present only in a minority of copies of the DNA in those cells where they are present, and if they are present also in only a minority of tumour cells then our power to detect them diminishes. On the other hand, as the quantity of starting DNA diminishes, sampling variation may lead to artificially low allele fractions in the sequencing library for some SNVs, and thus reduce our power to detect them.

Further to this, it is often not possible to obtain ‘pure’ tumour in the sense of all cells present in the sample being malignant. Indeed, the majority of cells in a sample may not be tumour cells, diluting the expected allele fraction of any mutations even further. We should note that the need to obtain ‘high cellularity’ samples in order that this effect doesn’t become too disruptive, whether occurring at the point of sampling or as a selection step where only high-cellularity samples go forward for sequencing, could conceivably bias the phenotypic nature of the cancer (e.g. away from more invasive parts of the tumour where it would be harder to avoid benign tissue, or avoiding substantial vascular growth, or simply that phenotypic differences affect the estimation of cellularity content) and therefore potentially also biasing the genomics.

We should note also that the normal sample may on occasion be contaminated with tumour material (esp if neighbouring ‘benign’ tissue is used or if the cancer studied is of the blood).

Issues 2: The sequencing library as representation of the sample presented:

Whatever the resolution to our sampling, having obtained samples to represent our tumour and normal we must process them in order to get them onto the sequencer. There may be multiple steps in this process (extraction, possible amplification, possible purification, sampling of the DNA for library preparation, the library preparation itself, and sampling by the sequencer) none of which we have any basis other than optimism for assuming to be unbiased. For evidence that each lane of sequencing is indeed a biased sample from the library, see for example the PhD work of Mike Smith (formerly at partner 4 (UCAM), now at partner 1 (EMBL)) [2].

Even if the individual steps are unbiased, we have seen in the PhD work of Daniel Andrews [3] that high variability in an early sampling step can lead to the estimated allele fraction of a mutation being quite different to that actually present at the beginning, and that consequently the precision with which we can estimate the allele-fraction of an SNV is difficult to assess. It should be clear that any attempts to improve the precision of an allele fraction estimate by 'deep-sequencing' will be improving the estimate of the allele fraction at the last stage of sampling, and not necessarily improving the estimate of the 'true' allele fraction.

Issues 3: The sequencing as a representation of the library.

Despite performing sequencing of tumour and normal to a nominal depth of coverage, a number of factors limit the probability of seeing an SNV even if it is represented in the sequencing library. Coverage is variable with known biases (e.g. depth of coverage associated with GC content) that will likely be correlated between the tumour and normal, but will also have a stochastic element. Thus the location of a mutation may not see any sequencing in the tumour (meaning it cannot be called), or it will not be sequenced in the normal sample (meaning that it cannot be identified as somatic).

Moreover the sequencer makes errors. Typically the error profile worsens along the sequence (meaning that generally we have more certainty in earlier bases), and following particular, often repetitive, sequences. While an estimate of the accuracy of a call can be made by the sequencer, and this may be further calibrated when analysing the data, we will require more than minimal representation of an SNV before we believe it. Certainly we require more than one read to be sequenced supporting an SNV before we will accept it. Similarly, we may choose not to reject an SNV call on the basis of a single read in the Normal tissue.

Issues 4: The interpretation as a representation of the sequencing.

In most cases, the sequence will be aligned to a reference genome in order to call mutations, although new classes of reference-free mutation calling software are appearing. The process is therefore reliant on the aligner performing well. There are essentially two reasons why aligners may perform poorly - too many alignment options or no option. The former occurs because of sequences that are not unique within the genome (at least to within the degree of uncertainty caused by potential sequencing error) and the latter because either there were errors in assembling the reference genome, or locally the reference genome is not a good match for the organism being sequenced.

Calling + Filters

In general, the calling procedure is two-fold. The question of whether or not two allele fractions differ is a well-defined question of probability, and differences in approach will depend on how the implementer wanted to model noise, what assumptions she wanted to make about allele fractions, and what external information to bring in (be that an external data set, or data from other loci in the same data set).

Subsequent to this, a set of filters are applied that account for various biases in the sequencing. Several tools have these in-built, but recently published advice also recommends “Filter[ing] by mapping quality, strand bias, positional bias, presence of soft-clipping to minimize mapping artefacts” [1].

Approaches to representing performance

Theory

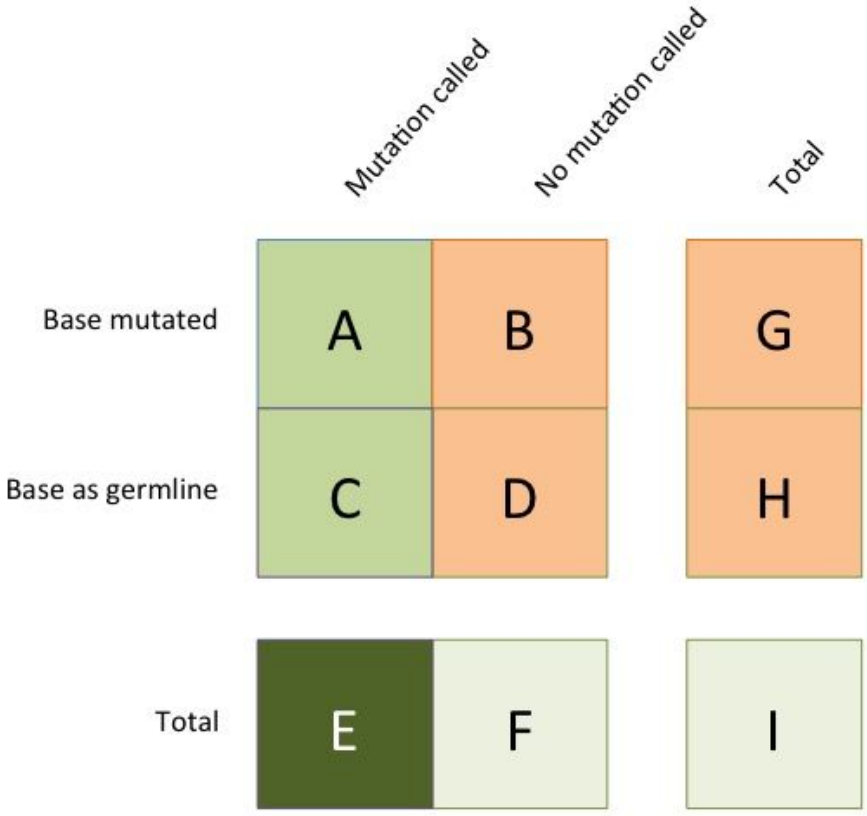


Figure 1. Representation of test performance with colour indicating whether the value is known to us.

The problem we have is that in general we only know the number of mutations that have been called (E in Figure 1). With some form of verification experiment, we can identify the breakdown of E into A and C.

Remarkably, the total number of bases (I), and therefore the number of bases that have not been called as mutated, is not strictly known. Should it be the size of the reference genome, the actual size of the patient's genome, the size of the mappable genome (albeit this is also not precisely defined), or the proportion of the genome with adequate coverage? Fortunately, given a choice of the above, the number tends to be of a magnitude that imprecision won't matter.

The problem comes in identifying B and D. The value of F will tend to be very large, while the value of B will be small relative to F (rendering it expensive even to estimate), but potentially large relative to A.

Representations of performance tend to focus on statistics such as sensitivity/specificity and precision/recall.

Sensitivity in this case is defined as A/(A+B) = A/G (also 1-FNR)
Specificity in this case is defined as D/(C+D) = D/H (also 1-FPR)

Precision in this case is defined as A/(A+C) = A/E (1-FDR)
Recall in this case is defined as A/(A+B*) = A/G* (=Sensitivity)

The attraction of precision as the only measure that we are actually in a position to estimate is obvious. However it should be noted that there are drawbacks. While sensitivity is estimated solely within those bases that are mutated and so (given certain assumptions) estimates of sensitivity may be thought to be transferable from one experiment to another carried out under the same conditions (the same for specificity), precision is dependent on the ratio of G to H and will thus vary dependent on the number of SNVs truly present.

Moreover it should be noted that if validating calls on only a small number of samples to represent a cohort, then one can manipulate performance by careful selection of samples.

Nevertheless, it is feasible that precision could be estimated for all samples in a study, and so we consider approaches for estimating recall in each case.

Practice

We consider the approaches of all of the primary (single-cancer) published TCGA studies (at time of finishing report) in representing uncertainty in their SNV calling and their recurrently mutated gene calling. This is not in any way intended to single out TCGA for criticism, but rather is a recognition of the field-leading role that that project has taken. It is also a convenient externally defined set of manuscripts (https://cancergenome.nih.gov/publications) that track the discipline over the last decade. The manuscripts are summarized in table 1.

Year	Cancer	#Cases	#SNVs	Validation	#Genes	Uncertainty
2008 [4]	Glioblastoma	91	453 (NS)	Precision	8	FDR<0.001
2011 [5]	Ovarian	316	19356	Precision +	9	FDR<0.15

				Recall(a)		
2012 [6]	Colon and Rectal	276	90060		32	
2012 [7]	Lung (SC)	178	48690 (NS)	Precision + Recall(b)	10	FDR<0.1
2012 [8]	Breast	507	28319		35	FDR<0.05
2013 [9]	AML	200	2315	Precision	23	FDR<0.05
2013 [10]	Endometrial Carcinoma	248	181930	Precision (c)	37	FDR<0.02
2013 [11]	Clear Cell Renal	417	36353	Precision	19	FDR<0.1
2013 [12]	Glioblastoma	291	20448	Precision (d)	71	FDR<0.1
2014 [13]	Bladder	130	38012	Precision	32	FDR<0.1
2014 [14]	Lung (A)			Precision (e)	18	Corrected p<0.025
2014 [15]	Gastric (A)			Precision (f)		FDR<0.1
2014 [16]	RCC	66	?	?	?	?
2014 [17]	Thyroid	406	6454	Precision (g)	7	FDR<0.1
2015 [18]	Head and Neck	279	49220	Precision (h)	11	FDR<0.1
2015 [19]	Glioma	293	9885	Precision (i)	19	FDR<0.1
2015 [20]	Melanoma	318	228,987	Precision (d)	42 (Mutsig)	FDR<0.1
2015 [21]	Breast	817	8,173	?	68	FDR<0.1
2015 [22]	Renal	161	10,380	?	5	FDR<0.1
2015 [23]	Prostate	333	14045	?	13	FDR<0.1
2016 [24]	Glioma	1122	714,305		75	
2016 [25]	Adrenocortical Carcinoma	91	8814	Precision (j)	5	FDR ≤ 0.1
2017 [26]	Oesophagus	164	?	?	5	? (k)

Table 1: Summary of representations of uncertainty/confidence in the identification of SNVs and recurrently affected genes within primary papers of the TCGA project.

(a) Recall assessed by examination of TP53, 20/303 TP53 mutations were missed.
(b) Calls that failed filters were validated as being false
(c) No validation in hypermutated cases
(d) Validation only of calls in recurrently mutated genes

(e) Validated high-power calls only
(f) Limited validation
(g) Heavily weighted towards calls in recurrently mutated genes
(h) Targeted regions only
(i) Limited mutations in recurrent genes and Glioma-associated ones
(j) Targeted and 'control' regions, but investigations of recall not indicated
(k) q-values reported, but cut-off not clearly indicated.

Several aspects are clear from Table 1. First, the numbers of SNVs identified in different cancer types do indeed vary by orders of magnitude. Second, the focus of reporting uncertainty in SNV calling has been on providing data on precision (if not often explicitly portrayed in that manner). Third, in more recent papers there has been less focus on reporting this aspect of performance. We can speculate that this results from growing confidence in the tools being used as the field develops, but note as above that the performance for precision will depend not only on the performance of the tool, but also the nature of the study, and so relying on previous studies for confidence may not be providing the full picture of performance.

The fourth observation is that validation studies are conflicted between answering questions about general performance (generating meta-data), and questions of biological interest (generating data); *a priori*, a mutation in a gene that is frequently mutated in the given cancer type is more likely to be real than one elsewhere. In those cases where where the motives are not conflicted, experiments tend to be designed solely for the latter. This conflict is natural given the competing pressures on laboratories, but the 'validation' aspect of these experiments is consequently minimal. Observations are only validated if the concern the question of the moment, which reduces a little the value of the data as a resource for the wider community.

Fifth, the practice of identifying significantly mutated genes has superficially changed little across the time period, although the frequently-employed Mutsig programme will itself have changed over that time and ,while not apparent from the table, the sophistication of interpretation of FDR has improved over time. Perhaps most notable, is that the degree to which false discoveries can be tolerated seems to have standardized regardless of context. Not only are the numbers and natures of recurrently mutated genes vastly different between projects, but the purposes for which they are being identified differ also. It seems however that 0.1 is becoming as standard for FDR in this field as 0.05 has been for sizes of statistical test.

Desired performance

It is worth taking a moment to consider what we want to achieve in terms of performance at calling SNVs. Projects are often seeking to identify recurrent mutations. Indeed the ICGC

projects have a nominal power calculation based on this premise, and there are two ways to view this question.

If we are incorporating the design of the sequencing experiment as part of the decision, then achieving high (perfect?) recall and precision is likely to be at the expense of sample size. The trade off between reduced performance in the individual and increasing the size of the cohort will be complex and likely to come down in favour of relatively low recall and precision. The power to detect recurrently mutated genes is not the sole aim of such a study, and in such circumstances it is likely that the ability to identify structural variants, mutually exclusive SNVs etc. will need to be factored in.

If we have generated the sequencing data already, then naturally it would be desirable to be at 100% recall and 100% precision, however given that in reality there has to be a trade off between the two, where should that be?

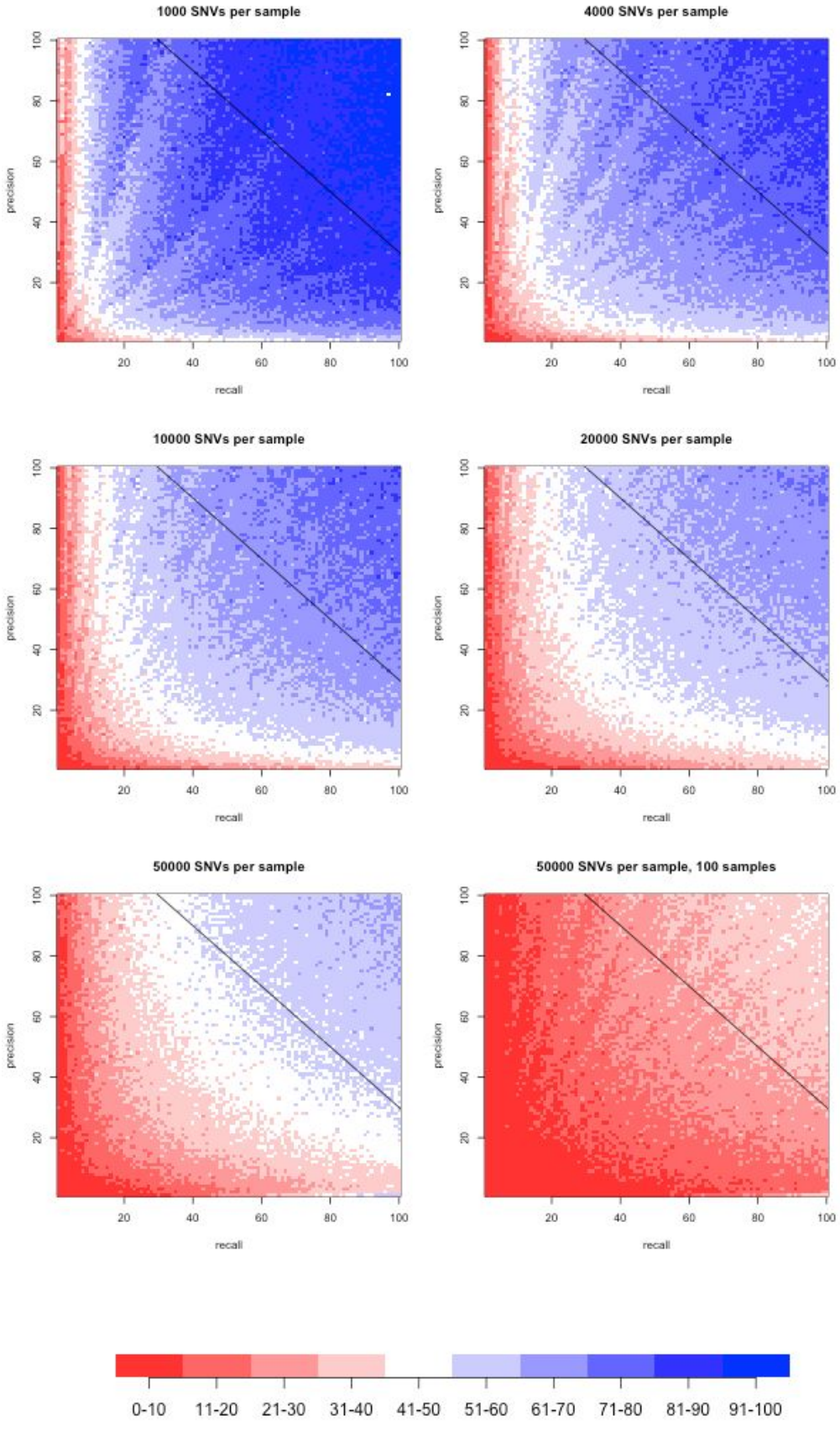
For illustration purposes, we don't use the power calculation associated with ICGC studies [27]. While a useful guide, this assumes that no false-positive results are tolerable and we have seen already that as many as 15% have been tolerated. We perform simulations, keeping the ICGC calculation's 20,000 genes of size 1,500 bases and 500 samples (unless indicated otherwise). We simulate 100 recurrently mutated genes with (arbitrary) population rates as given in Table 2 and ask what number of these end up in a list of the top 100 genes with most evidence of recurrence - a summary that is perhaps more in-keeping with practice. Details of the performance are given in Figure 2.

As would be anticipated, the performance is generally worse as the background mutation rate increases, and substantially worse when the number of samples decreases. Note that, in general, the effects of recall and precision are well-balanced on this scale and an approach of 'as close to the top-right as possible' will serve well. There is though always a bias towards sacrificing precision ahead of recall.

Population rate	75 %	50%	20%	10%	5%	2%	1%
Number	1	1	8	20	20	20	30

Table 2: Simulated population rates of 100

Figure 2: Charting for different SNV precisions and recalls, the ability in simulation to identify 100 recurrently mutated genes.



Approaches to estimating numbers of missed SNVs.

As noted, the estimation of B through the mechanism of sampling from F is unlikely to be efficient enough to be practicable. Sequencing the sample to high depth is the ideal, but to do this in every case would be expensive and nonsensical.

The SNVs that are missed fall into two categories: Those due to bias and those due to bad luck. We suggest that the numbers of SNVs missed due to chance are estimable, and that this may suffice as an estimate of B.

To take a trivial case suppose that we saw 800 SNVs, and could estimate for each SNV the power that we had had to detect each SNV (under some model), then if each SNV had a post-hoc power of 80%, then we might surmise that these 800 identified SNVs represent 1000 in total.

In the case of strong filters after calling (minimal supporting reads in the matched normal, minimum coverage, minimum support in the tumour), the critical region for the initial test is well approximated by n>k where n is the number of reads supporting a mutation in the tumour. In theory we could also account for any strand bias and positional bias filters that are applied.

The software package as part of this deliverable is an initial implementation of the approach to estimate B and allow for further analysis as discussed below [28]. This prototype has been developed by Theresa Schüler (under the supervision of Andy Lynch) and Daniele Biasci. A mature version will emerge from collaboration with partners on the SOUND project.

Meaningful interpretation

As has been noted, there is heterogeneity in the set of SNVs. Some may be supported by hundreds of reads with others supported by only three. To summarize the set of called SNVs by a single (or pair of) statistic(s) such as precision (or precision/recall) is simplistic, and masks the fact that we have greater confidence in some than others.

Recall that we are motivated by finding recurrently mutated genes in a population. Were we to enforce the same precision in each sample, then there would be instances of SNVs presenting with equal evidence in different samples and achieving different outcomes in terms of being

called. If as is commonly the case, the level of evidence is required to be consistent, then a single measure of precision may not suffice.

An overall precision estimate would be dominated by the performance of any hypermutated samples present, while the average of the sample-specific precisions may be better, but could tell us nothing about an individual case. We would suggest that presenting the distribution of precisions (and recalls or equivalent) would be a better option for such a study.

Experimental design

The software package we present as part of deliverable 11.1 offers a natural extension that would be useful for experimental design. From estimating the number of SNVs that have been missed (B), it is not too difficult to estimate the number of additional SNVs that a further 10x (for example) sequencing coverage would reveal.

As well as aiding design, this might provide a more interpretable alternative to recall (which we have had to define in a slightly obscure manner) as a summary of performance. It could also aid allocation of resources in any sequencing initiative and allow for two-phase or even fully-adaptive experimental designs.

A further extension is the characterisation of the missed SNVs. We can not only estimate the total number missing, but also within categories (e.g. mutation types, genomic locations, age relative to copy number etc.). The characterization of that which we are missing can therefore facilitate the design of validation experiments, or at least allow for a weighted interpretation of those experiments if not allowed to contribute to the design.

References

[1] Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., … Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*, 10001.

[2] Smith M. L. (2013) Low-level artefacts affecting microarray and next-generation sequencing data in a cancer genomics environment. *PhD Thesis*, University of Cambridge.

[3] Andrews D. J. (2015) Statistical models of PCR for quantication of target DNA by sequencing. *PhD Thesis*, University of Cambridge.

[4] McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M. Mastrogianakis, G., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216), 1061–1068.

[5] Cancer Genome Atlas Research Network, Bell, D., Berchuck, a., Birrer, M., Chien, J., Cramer, D. W., … Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609–615.

[6] Cancer Genome Atlas Network, Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. a., … Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337.

[7] Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., … Meyerson, M. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*(7417), 519–525.

[8] Cancer Genome Atlas Research Network, Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., … Tarnuzzer, R. W. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *487*(7407), 61–70.

[9] Voigt, P., & Reinberg, D. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia The Cancer Genome Atlas Research Network. *The New England Journal of Medicine*, *368*(22), 2059–74.

[10] Getz, G., Gabriel, S. B., Cibulskis, K., Lander, E., Sivachenko, A., Sougnez, C., … Levine, D. a. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73.

[11] Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. a., Gibbs, R. a., Gordon Robertson, A., … Sofia., H. J. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, *499*(7456), 43–49.

[12] Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., … McLendon, R. (2013). The somatic genomic landscape of glioblastoma. *Cell*, *155*(2), 462–477.

[13] Cancer Genome Atlas Research Network, Weinstein, J. N., Akbani, R., Broom, B. M., Wang, W., … Eley, G. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, *507*(7492), 315–22.

[14] Eric, A. C., Joshua, D. C., Angela, N. B., Alice, H. B., William, L., Juliann, C., … Ming-Sound, T. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, *511*(7511), 543–550.

[15] Cancer Genome Atlas Research Network, Bass, A. J., Thorsson, V. V., Shmulevich, I., Reynolds, S. M., … Liu, J. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202–209.

[16] Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., … Creighton, C. J. (2014). The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma. *Cancer Cell*, *26*(3), 319–330.

[17] Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., … Zou, L. (2014). Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell*, *159*(3), 676–690.

[18] Lawrence, M. S. M. S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Gabriel, S. B. S. B., … Pham, M. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, *517*(7536), 576–582.

[19] Cancer Genome Atlas Research Network, Brat, D. J., Verhaak, R. G. W., Aldape, K. D., Yung, W. K. A., Salama, S. R., … Zhang, J. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England Journal of Medicine*, *372*(26), 2481–98.

[20] Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., … Zou, L. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681–1696.

[21] Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., … Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506–519.

[22] The Cancer Genome Atlas Research Network. (2016). Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *The New England Journal of Medicine*, *374*(2), 135–45.

[23] Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., … Zmuda, E. (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, *163*(4), 1011–1025.

[24] Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., … Verhaak, R. G. W. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, *164*(3), 550–563.

[25] Zheng, S., Cherniack, A. D., Dewal, N., Moffitt, R. A., Danilova, L., Murray, B. A., … Verhaak, R. G. W. (2016). Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell*, *29*(5), 723–736.

[26] Kim, J., Bowlby, R., Mungall, A. J., Robertson, A. G., Odze, R. D., Cherniack, A. D., … Zhang, J. (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature*.

[27] icgc.org/index.php?q=icgc/goals-structure-policies-guidelines/e7-study-design-and-statistical-issues

[28] https://github.com/db544/varmiss