Accurate genotyping of *UGT1A1* dinucleotide repeat polymorphism from targeted NGS data for the assessment of irinotecan chemotherapy adverse events

Francisco M. De La Vega[1], Len Trigg[2], Kurt Gaastra[2], Sean A. Irvine[2], Gene Selkov[1], Yan Yang[1], Kyung Choi[1], and Robert Huether[1]

[1]Tempus Labs, Chicago, IL, USA; [2]Real Time Genomics, Ltd., Hamilton, New Zealand

INTRODUCTION

Colorectal cancer (CRC) is a leading cause of cancer-related death across the world. Irinotecan (IRI) is commonly used to treat metastatic CRC. The gene *UGT1A1* encodes the enzyme responsible for the glucuronidation of SN-38, the active metabolite of IRI. The TA repeat in the promoter region of *UGT1A1* is highly polymorphic. Wild-type *UGT1A1* contains six TA repeats [A(TA)$_6$TAA]. Polymorphic *UGT1A1* alleles with a higher number of TA repeats, such as UGT1A1 *28 /(TA)$_7$ and *37/(TA)$_8$ alleles, decrease promoter activity and are associated with severe toxicity in patients receiving IRI-based chemotherapy, for which dose reductions are recommended. Matched tumor/normal genomic profiling by NGS for cancer therapy may be useful to assess therapy-induced adverse events due to germline variants such as those in *UGT1A1*. However, genotyping of *UGT1A1* polymorphisms is commonly carried out with PCR or fragment analysis in capillary electrophoresis, and not from NGS data. This is due to challenges in aligning short reads to repeats and the introduction of "stutter" artifacts due to DNA polymerase slippage that add or delete copies of the repeat unit in the observed sequencing reads[1]. Herein, we benchmark a novel method, *BayeSTR*, to call accurate *UGT1A1* TA repeat genotypes from target capture NGS data and demonstrate the feasibility of this method for genomic profiling of cancer patients.

METHODS

BayeSTR analyzes deduplicated read alignments to a graph-based model representing the possible repeat alleles[2], and then performs genotype calling by a Bayesian model that incorporates an empirically derived DNA polymerase stutter[1] denoising model. The Bayesian model provides genotype posterior probabilities as confidence values that can be used to eliminate genotyping errors for poor quality data/samples.

Figure 1. UGT1A1 analysis workflow

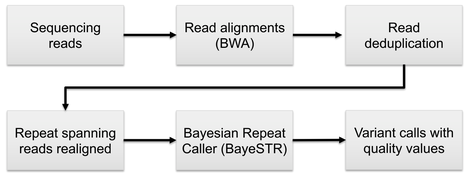
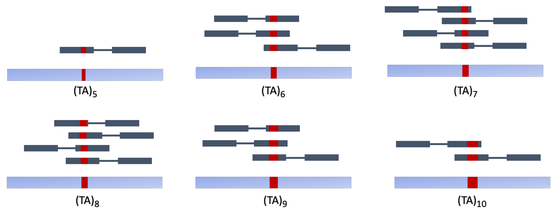


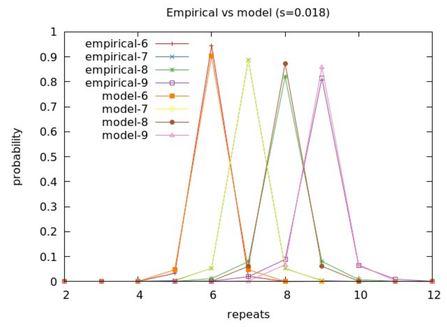
Figure 2. Realignment of repeat spanning reads



Repeat-spanning read alignments produced by BWA are de-duplicated and realigned locally to several models of the reference including the different repeat lengths expected[2]. These alignments provide read counts that are used by BayeSTR to test hypothesis of possible genotypes using a stutter model derived from empirical observations in 1,419 patient blood samples sequenced with the 648-gene Tempus xT NGS assay (see Fig 3).

METHODS

Figure 3. Stutter spectrum error model



The empirically observed stutter distribution is well approximated by a simple one-parameter (probability of TA insertion, s) model. The model can generalize to arbitrary repeat lengths for which empirical data was not available. This model is used to construct priors when evaluating genotype hypothesis by BayeSTR (see Table 1).

Table 1. Example – from read counts to genotype call

Repeat Length	Read Count
5	2
6	16
7	140
8	116
9	23
10	3

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Hypothesis Scored				
Allele A	Allele B	log posterior prob	log odds ratio	GQ
5	5	-2930.28		
5	6	-1737.44		
5	7	-782.39		
5	8	-746.62		
5	9	-1391.93		
5	10	-1998.13		
6	6	-1539.04		
6	7	-741.68		
6	8	-624.57		
6	9	-956.98		
6	10	-1343.19		
7	7	-596.34		
7	8	-365.37	230.9	1003
7	9	-538.94		
7	10	-675.33		
8	8	-610.47		
8	9	-730.81		
8	10	-781.31		
9	9	-1356.24		
9	10	-1545.22		
10	10	-2335.41		

Simulations

To benchmark our method, we simulated alignment data for TA repeat lengths from 4-10 copies in multiple genotype combinations and coverage depths using data from 224 patient blood samples sequenced with the 648-gene Tempus xT NGS assay (deduplicated coverage ranging 200-500x). This is constructed by partitioning the alignments present in the full set of available samples based on the alignment CIGAR. From these available reads it is possible to simulate samples containing alleles not actually expected to be seen in the population (e.g. (TA)$_5$ or (TA)$_{10}$), to study the behavior of the genotype calling algorithm.

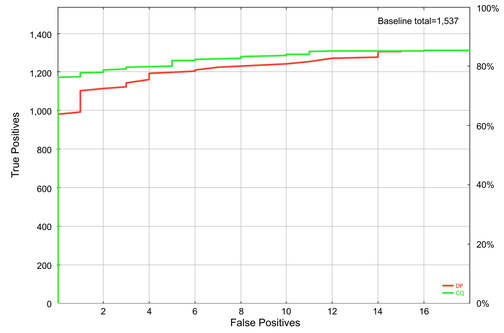
RESULTS

Table 2. Performance with simulated data

Truth	Cov	Call	GQ	Cov	Call	GQ
5/5	100	5/5	285	500	5/5	1242
6/6	100	6/6	218	500	6/6	1082
7/7	100	7/7	206	500	7/7	1039
8/8	100	8/8	203	500	8/8	990
9/9	100	9/9	192	500	9/9	956
10/10	100	10/10	273	500	10/10	1349
5/7	100	5/7	539	500	5/7	2618
6/7	100	6/7	342	500	6/7	1673
7/8	100	7/8	324	500	7/8	1651
7/9	100	7/9	472	500	7/9	2346
7/10	100	7/10	501	500	7/10	2500
6/9	100	6/9	493	500	6/9	2466
5/10	100	5/10	588	500	5/10	2812

We simulated several repeat allele lengths in different genotype combinations ("Truth") at both 100X and 500X depths of coverage ("Cov"). Genotype calls made ("Call") and genotype quality score ("GQ") for both coverages are shown, demonstrating the ability to correctly call known alleles as well as new potential alleles without misclassification.

Figure 4. ROC comparing variant quality scores



Receiver-operator curve showing the discriminating ability of two possible quality scores: read depth (DP, red) and Phred-scaled genotype quality (GQ, green). Based on these results we selected a GQ of 70 to filter raw variant calls and remove false positives.

Limit of Detection by Coverage

We observed that with a minimum depth of 70X, we obtained 100% accuracy and robustness to rare/new repeat alleles. We validated our method with germline data from the Tempus xT tumor-normal matched NGS test, which targets 648 cancer related genes including *UGT1A1*.

Table 3. Accuracy vs depth of coverage

Cov target	Samples	Calls	Match	Mismatch
10	224	8	5	3
30	224	223	210	13
50	224	224	223	1
70	224	224	224	0
90	224	224	224	0
100	224	224	224	0
200	220	220	220	0

We explored the dependency of the algorithm's ability to call repeat alleles on the depth of coverage. We tested it on a set of 224 patient samples sequenced with Tempus xT NGS test; the UGT1A1 repeat alleles in those samples were determined by an orthogonal method that searches for patterns in unaligned reads ("the silver set"). By subsampling the data, we simulated several levels of read depth. Our results suggest that a minimum of 70x is necessary for accurate results.

Validation

We sequenced DNA from 51 cell-lines from the Coriell Institute which are part of the CDC GeT-RM project with orthogonally confirmed *UGT1A1* genotypes. These samples include different combinations of 6-9 TA repeats. Additionally, we are currently evaluating performance with 66 patient samples which are being confirmed by orthogonally fragment analysis.

Table 4. Performance with Get-RM cell lines

				BayeSTR		Expansion Hunter	
Genotype	Repeat	Zygosity	Cell lines	Match	Mismatch	Match	Mismatch
*1/*1	7/7	Hom	15	15	0	15	0
*1/*36	6/7	Het	4	4	0	-	-
*1/*37	7/9	Het	2	2	0	-	-
*1/*28	7/8	Het	15	15	0	15	0
*28/*28	8/8	Hom	11	10	1	1	10
*28/*36	6/8	Het	3	3	0	-	-
*28/*37	8/9	Het	1	1	0	-	-

*We sequenced 51 reference Coriell cell lines previously characterized by the CDC Get-RM project[3] with orthogonally validated UGT1A1 repeat alleles. We compared calls by BayeSTR with those made by the Expansion Hunter[2] software. BayeSTR calls matched the truth set, except for NA20509, where a SNV is also present in the repeat. By contrast, Expansion Hunter exhibits a significant error rate for *28/*28 homozygotes. Data from cell-lines with *36 and *37 genotypes were not evaluated with Expansion Hunter.*

CONCLUSIONS

- *BayeSTR* allows for **automated and accurate *UGT1A1* promotor genotyping** from targeted NGS data and can be applied to other genomic repeat regions of clinical relevance.
- This method **identifies *UGT1A1* repeat polymorphisms associated with IRI-induced adverse events** and can be used during clinical NGS testing to further support clinician treatment decisions for cancer patients.

REFERENCES

[1]Raz, O. *et al.* Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Res* **47**, gky1318-(2019).
[2]Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions. *Bioinformatics* 35, 4754–4756 (2019).
[3]Pratt, V. M. *et al.* Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes A GeT-RM Collaborative Project. *J Mol Diagnostics* 18, 109–123 (2016).

TEMPUS