Cancer Analysis Workflow Sprint Review

SciLifeLab

14th November 2016

Cancer Analysis Workflow Sprint Review

What happened in the last month?

- Features in release v0.9
- Refactoring code
- DREAM challenge results
- Next steps

Although focus was on processing the DREAM challenge data, many new features were added, bugs fixed.

Features in release v0.9

This release is capable to process tumour/normal pairs starting from raw FASTQ files, or from preprocessed BAM files.

Features we are expected to work

- Preprocessing (Alignment, Merging... all the steps to get a base-recalibrated BAM)
- SNP and indel callers: MuTect1 and Strelka working
- SV caller: Manta

Features for better user experience

- Start from realignment: we have to realign BAMs together
- One can chose different targets (Manta,MuTect1,... and their combination)
- Restart only from the finished preprocessing step
- More documentation

DREAM challenge synthetic sets

First set of somatic call challenges

Using pre-processed BAM files provided by TCGA compare variant callers for better somatic call results.
Published results after several steps of optimizations
Our results are listed as the very first approach

- S1: 3537 somatic SNVs, 100% tumour, few indels and SVs
- S2: 4332 somatic SNVs, 80% tumour, some indels and SVs
- S3: 7903 somatic SNVs, 33% tumour, 20% different subclone, many indels and SVs

Getting consensus calls

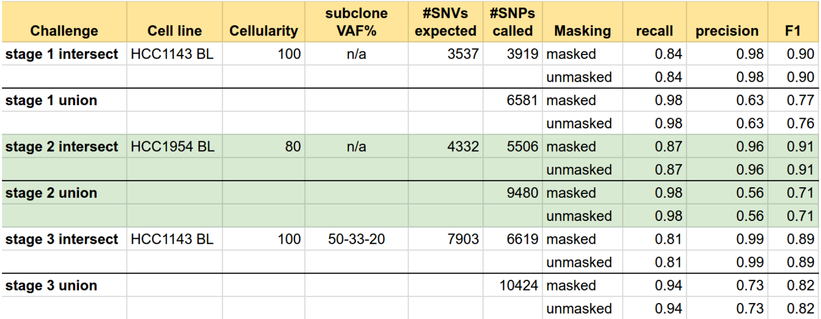
SNPs are called only if both MuTect1 and Strelka gives a call

- Expected to have a SOMATIC info flag (both caller classified the SNP as a somatic one)
- Expected to have a PASS filter by both callers (both caller was confident in the call)

Recall and precision was calculated by the python script provided by the ICGC-TCGA DREAM group

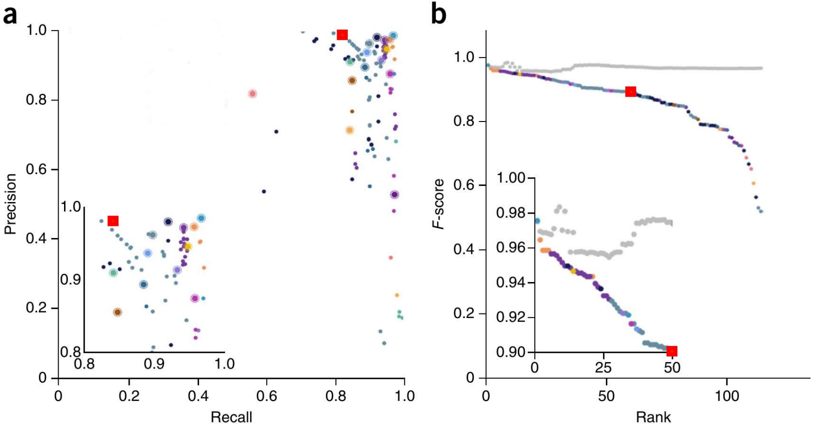
- $recall = \frac{TP}{TP+FN}$
- $precision = 1 - \frac{FP}{(TP+FP)}$
- $F_{score} = 2 \times \frac{precision \times recall}{precision + recall}$

DREAM challenge - SNPs

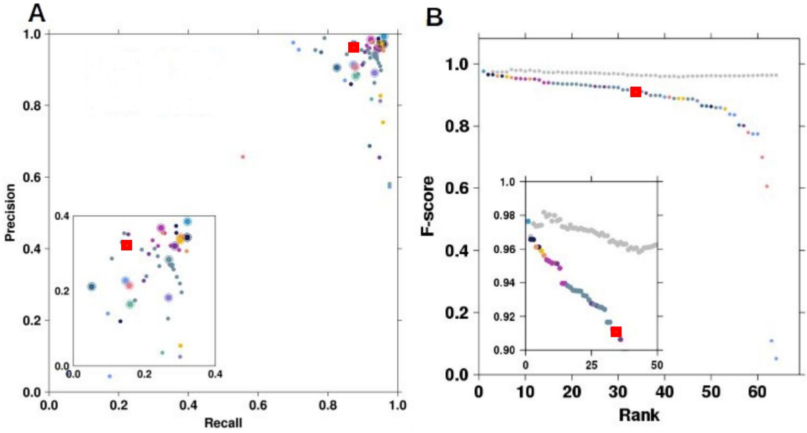


Challenge	Cell line	Cellularity	subclone VAF%	#SNVs expected	#SNPs called	Masking	recall	precision	F1
stage 1 intersect	HCC1143 BL	100	n/a	3537	3919	masked	0.84	0.98	0.90
						unmasked	0.84	0.98	0.90
stage 1 union					6581	masked	0.98	0.63	0.77
						unmasked	0.98	0.63	0.76
stage 2 intersect	HCC1954 BL	80	n/a	4332	5506	masked	0.87	0.96	0.91
						unmasked	0.87	0.96	0.91
stage 2 union					9480	masked	0.98	0.56	0.71
						unmasked	0.98	0.56	0.71
stage 3 intersect	HCC1143 BL	100	50-33-20	7903	6619	masked	0.81	0.99	0.89
						unmasked	0.81	0.99	0.89
stage 3 union					10424	masked	0.94	0.73	0.82
						unmasked	0.94	0.73	0.82

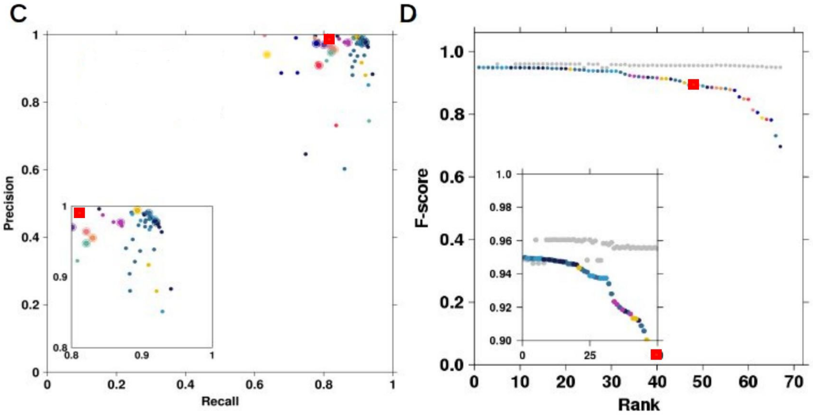
DREAM challenge - SNPs stage 1



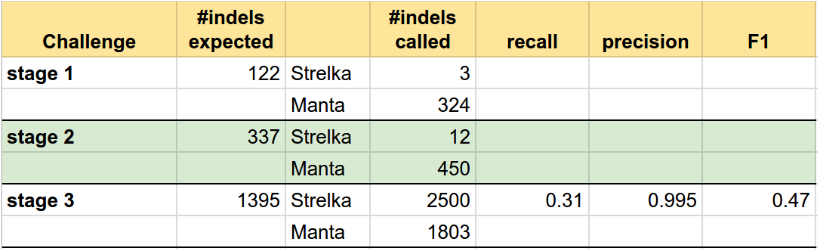
DREAM challenge - SNPs stage 2



DREAM challenge - SNPs stage 3

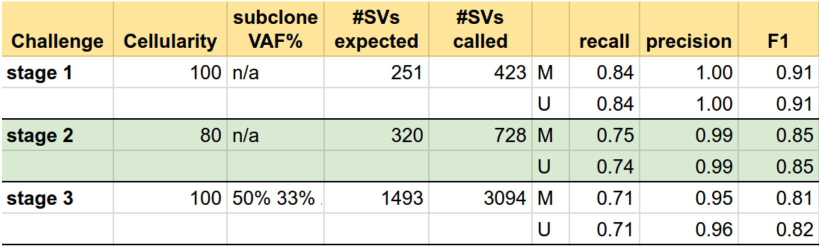


DREAM challenge - indels



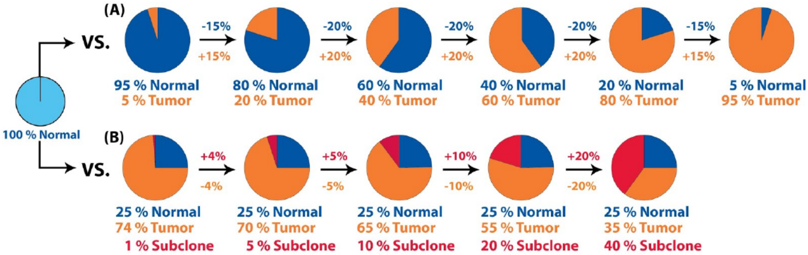
Challenge	#indels expected		#indels called	recall	precision	F1
stage 1	122	Strelka	3			
		Manta	324			
stage 2	337	Strelka	12			
		Manta	450			
stage 3	1395	Strelka	2500	0.31	0.995	0.47
		Manta	1803			

DREAM challenge - SVs

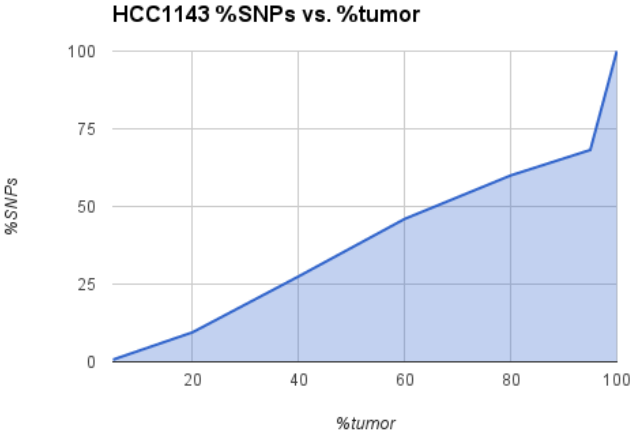


Challenge	Cellularity	subclone VAF%	#SVs expected	#SVs called		recall	precision	F1
stage 1	100	n/a	251	423	M	0.84	1.00	0.91
					U	0.84	1.00	0.91
stage 2	80	n/a	320	728	M	0.75	0.99	0.85
					U	0.74	0.99	0.85
stage 3	100	50% 33%	1493	3094	M	0.71	0.95	0.81
					U	0.71	0.96	0.82

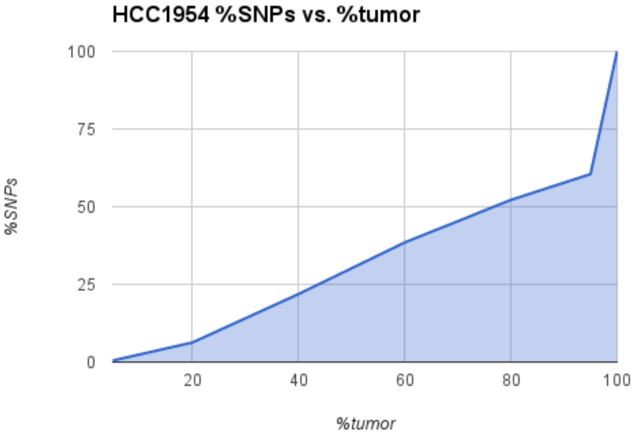
Purity and clonality tests in the DREAM challenge



Purity HCC1143



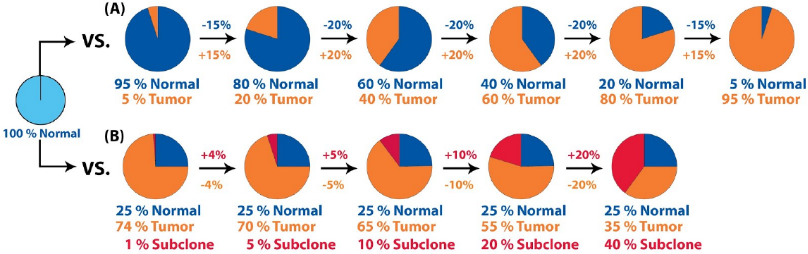
Purity HCC1954



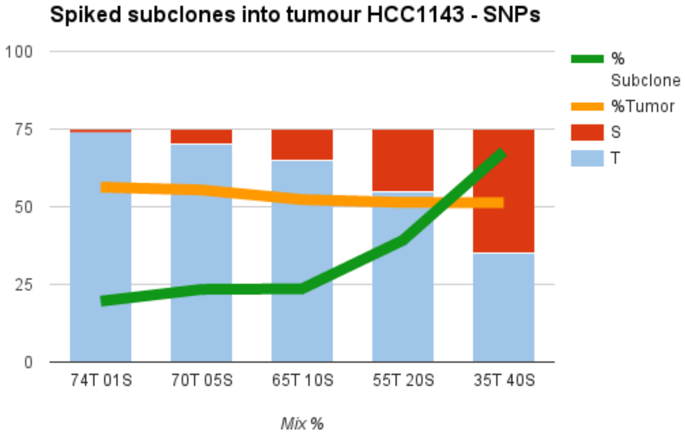
Conclusions about purity

- We are getting something we were expected to get
- No idea why we are loosing SNVs so quickly
- Sensitivity: at 20% tumour content we can see almost nothing (contradicts to challenge S3)
- More/better tests are needed
- Will be worth to compare results from other software (i.e. ASCAT)

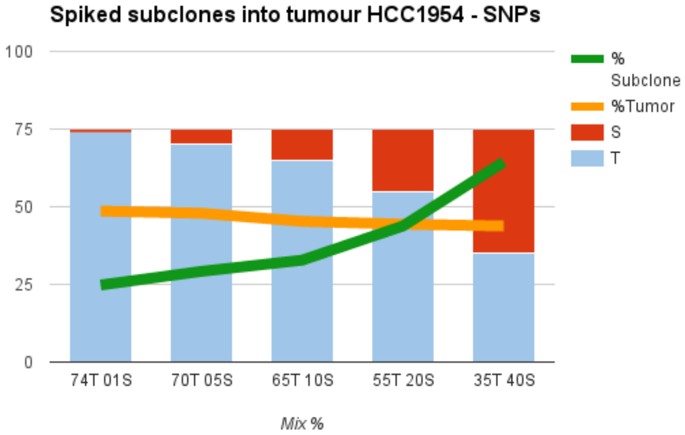
Purity and clonality tests in the DREAM challenge



Clonality tests HCC1143



Clonality tests HCC1954



Conclusions about clonality

- Not clear why we have high recall for subclones
- Seems we can have somatic calls at low tumour concentration
- Have no tools to distinguish clones (i.e. when adding relapse samples)

Towards version 1.0

- Refine SNV call recall *and* precision (get higher F values)
- Refine indels (look around for an alternative caller and improve selection)
- Add ASCAT
- Provide a merged VCF for ranking

Towards version 2.0

- Add more test cases (sensitive data: we have to fill out papers)
- Add CNVs (got relatively little focus)
- Persuade users to use the workflow: to surface bugs and features
- Refactoring: must happen continuously

Longer term plans

- Variant annotation
- Variant scoring/ranking
- Program cancer interface in Scout/Puzzle/New software
- Flexible choice in how to treat several variant callers (intersect, overlap, scoring etc)
- Set up CAW on Clinical Genomics' hardware
- Switch to GrCh38
- Exome/custom capture support and QC
- Integrate with RNA seq

Final remarks

- The workflow is robust - providing the underlying infrastructure works
    - We have no information about Bianca right now
- Have to be more user-friendly
    - Malin as primary test user on real data
    - Teresita already used the workflow, and found some trivial inconsistencies
    - Command-line interface is going to be stable
- Nextflow looks like a good choice, we have an active user/developer community
- I am actually surprised that the DREAM challenge results are OK - thanks for the input from Oslo!

TODO

- Fix the recall/precision rate values reported for Manta SVs
- Report SNPs/indels/SVs in a separate set of files
- Report all the four options for SNPs (MuTect1 only, Strelka only, union, intersection)
- Prioritize recall (sensitivity)
- Check Manta indel reports/possible optimization
- Control-FREEC / Canvas validation by Malin?