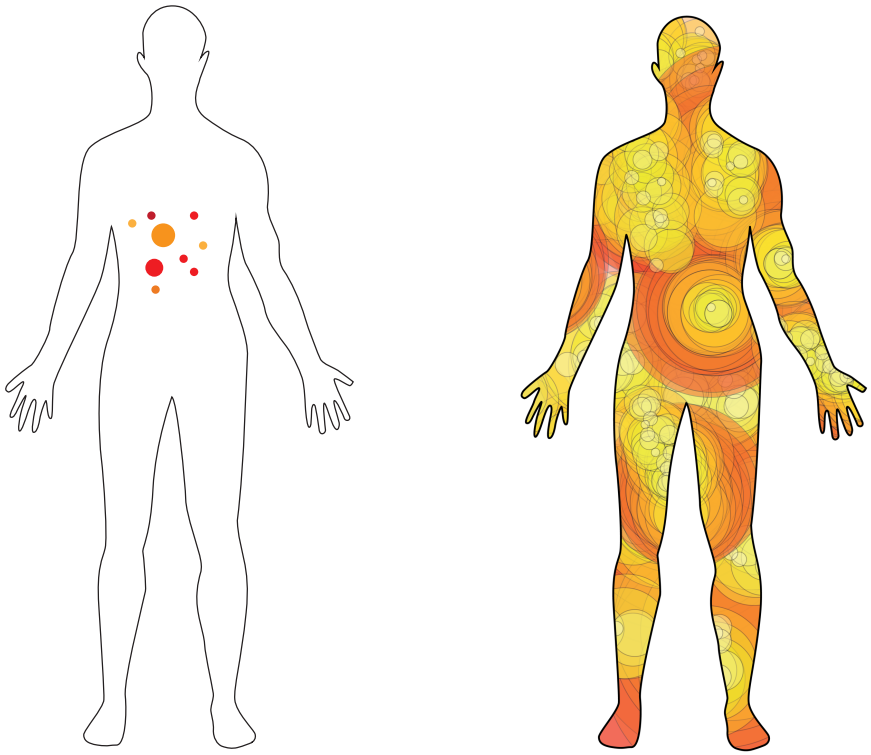
Juliet - One Click Minor Variant Calling

github.com/pacificbiosciences/minorseq

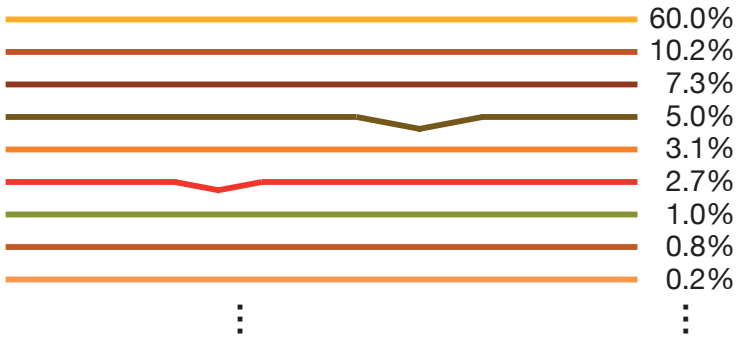
Dr. Armin Töpfer

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2017 by Pacific Biosciences of California, Inc. All rights reserved.

WITHIN-PATIENT VIRAL DIVERSITY



Infection Day 1 → After Months

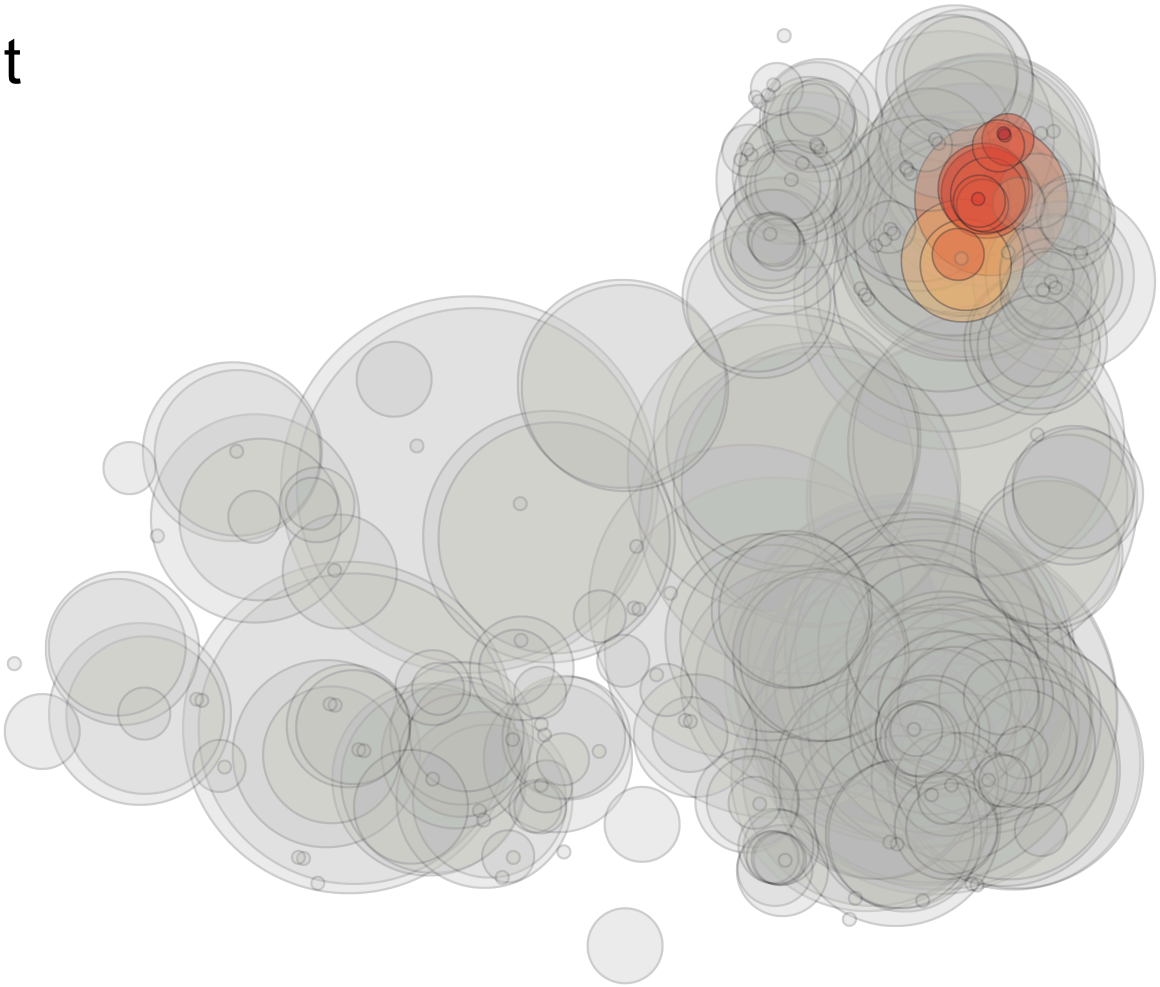


10^{10}-10^{12} virions per day

$3x10^{-5}$ mutations/RT/base

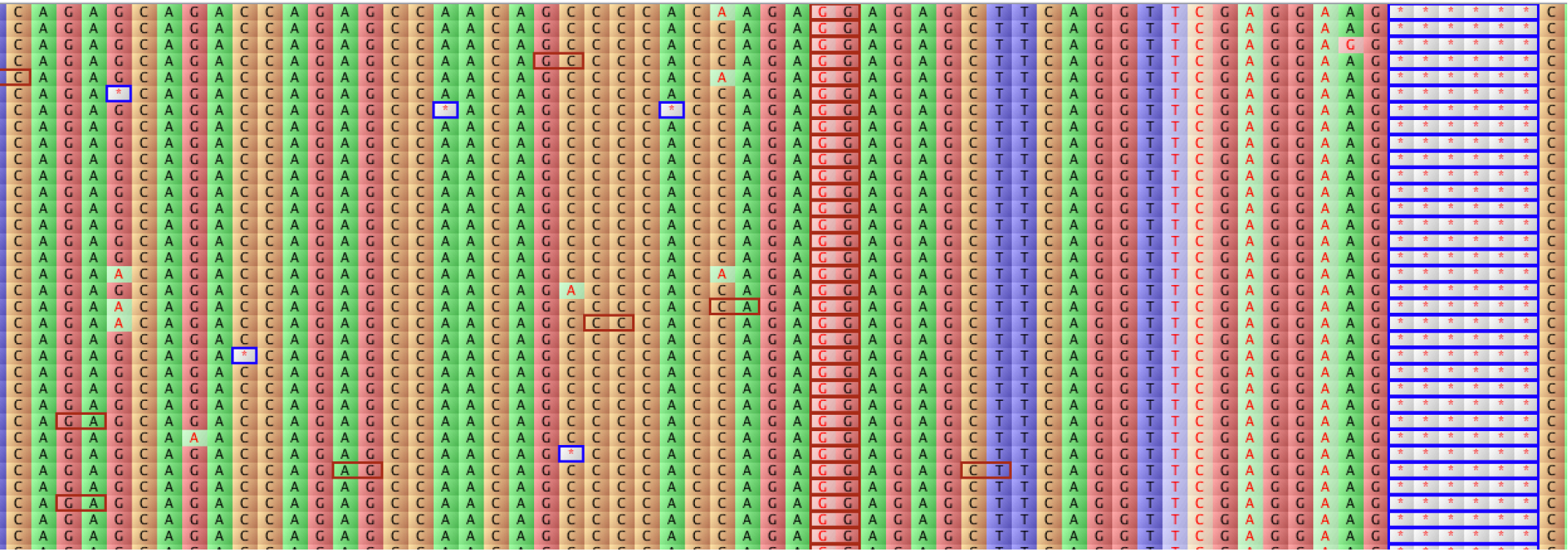
HOW TO AVOID TREATMENT FAILURES?

- **Identify** minor variants
- **Adjust** treatment
- **Avert** failure



SCALES OF RECONSTRUCTION

CHALLENGE



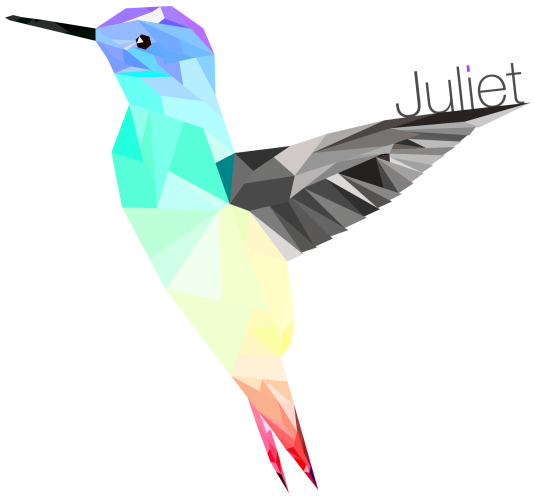
Reliable identify 1% variants from sequencing noise

JULIET

Reliable and reproducible 1% amino acid variant discovery tool

MINOR VARIANT CALLING - JULIET

- **Targeted amplicon approach, reference guided, one-click analysis**
  - <=4kb amplicon (fully spanned by long reads)
  - Initial viral focus (HIV, HCV, HBV), oncology BCR-ABL
  - De-novo codon variant discovery now, small in/dels later
- **Reliable 1% minor variant detection**
  - 10^{-6} false-negative rate & <=1% false-positive rate
  - 6000 high-quality CCS per sample (RQ > 0.99, ≳ 5 passes)
- **High multiplexing per chip, 6 hour run**
  - 8 samples for 1% accuracy: 48k CCS reads yield
  - >50* samples for 10% accuracy: 30k CCS reads yield
- **Extensible for new disease areas and organisms**
  - Drug-resistance mutation annotation
- **Alpha test in progress – MSSM and Uppsala**



WHY 6000X FOR 1% @3200BP?

Minor Percentage	Minimal Coverage	Reliable* Coverage
1%	2500	6000
5%	500	1200
10%	250	600

*for 1 FN per million samples

STATISTICAL METHOD: FISHER'S EXACT TEST

Compute probability that the observed codon frequency corresponds to the expected CCS noise:

1) One-sided Fisher's exact test:

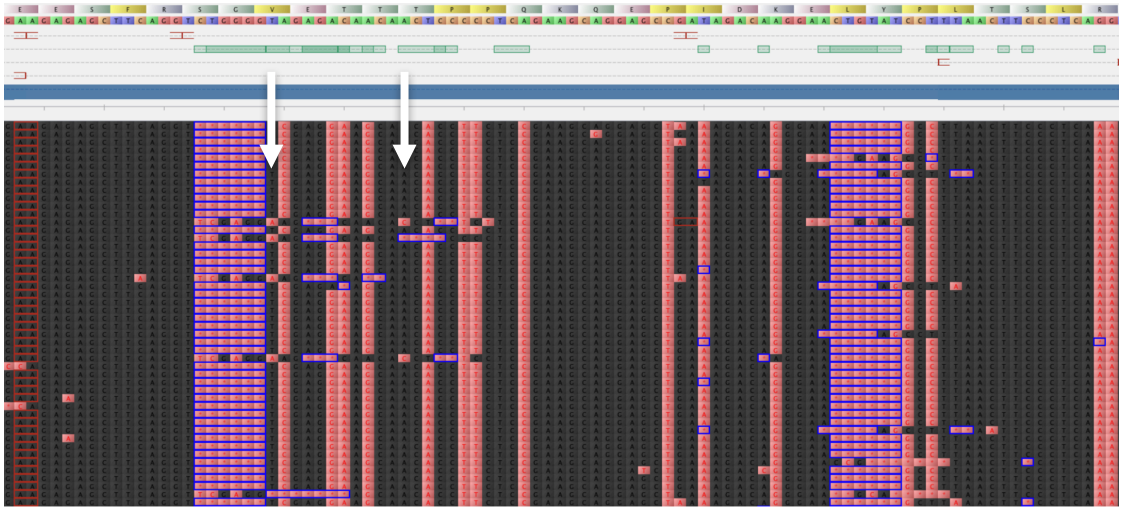
```
p = FET( observed_counts, coverage-observed_counts,
         expected_counts, coverage-expected_counts)
```

2) Bonferroni correction.

3) Call codon to be significant if $p < \alpha$.

REMAPPING QUALITY

Naïve alignment against HXB2



mismapped deletions

REMAPPING QUALITY

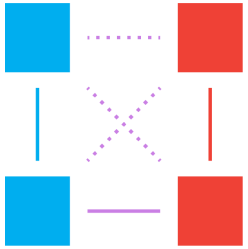
Naïve alignment against HXB2

Re-alignment strategy against HXB2

UNIFIED MINOR VARIANT SOFTWARE PIPELINE - JULIETFLOW

BLASR

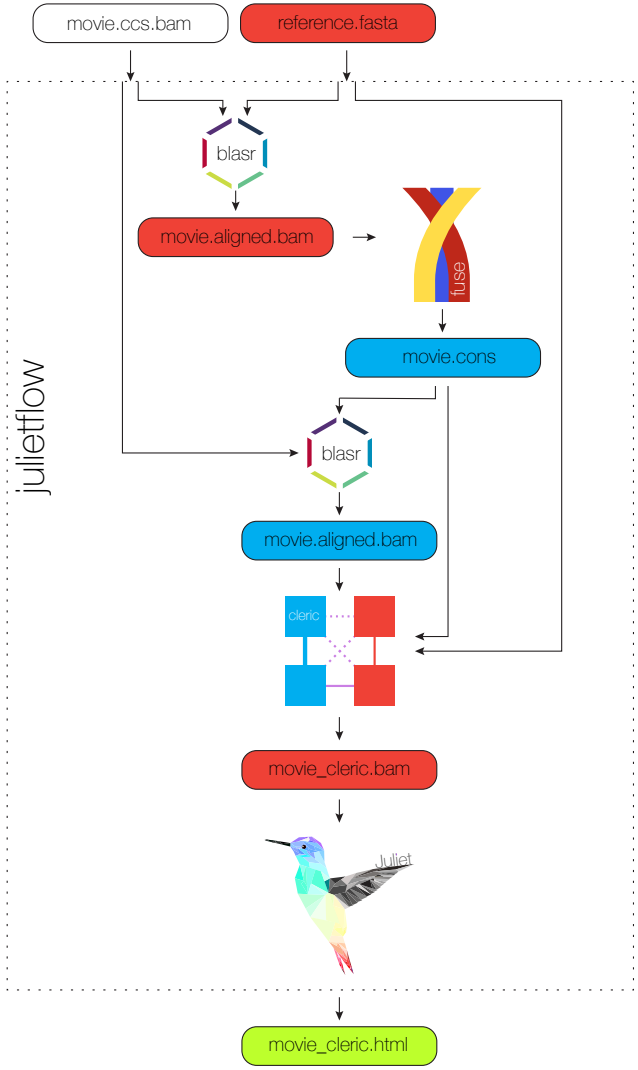
fuse



cleric

juliet

UNIFIED MINOR VARIANT SOFTWARE PIPELINE - JULIETFLOW




```
$ julietflow                \
    -i movie.ccs.bam        \
    -r reference.fasta
```

JULIET OUTPUT

Output is JSON and its HTML representation.
Directly interpretable, incl. drug-resistance mutation annotation

Protease							
HXB2			Sample				
Codon	AA	Pos	AA	Codon	Frequency	Coverage	DRM
GTC	V	3	I	ATC	0.97	6001	
AGT	S	37	N	AAT	0.9	6001	

Reverse Transcriptase							
HXB2			Sample				
Codon	AA	Pos	AA	Codon	Frequency	Coverage	DRM
ATG	M	41	L	TTG	0.0094	6001	NRTI S
AAA	K	65	R	AGA	0.0096	6001	NNRTI S + NRTI S
AAG	K	102	Q	CAG	0.95	6001	
GAA	E	122	K	AAA	0.9	6001	
AGT	S	162	C	TGT	0.96	6001	
TAT	Y	181	C	TGT	0.0096	6001	NNRTI + NRTI
GGA	G	190	A	GCA	0.0099	6001	NNRTI + NRTI
CTT	L	214	F	TTT	0.91	6001	


```
{
  "genes": [
    {
      "name": "Reverse Transcriptase",
      "variant_positions": [
        {
          "coverage": 6001,
          "ref_amino_acid": "M",
          "ref_codon": "ATG",
          "ref_position": 41,
          "variant_amino_acids": [
            {
              "amino_acid": "L",
              "variant_codons": [
                {
                  "codon": "TTG",
                  "frequency": 0.00949841693051158,
                  "known_drm": "NRTI S",
                  "pValue": 4.00758706533135e-13
                }
              ]
            }
          ]
        },…
```

JULIET OUTPUT

Output is JSON and its HTML representation.
Directly interpretable, incl. drug-resistance mutation annotation

Protease

HXB2 Codon	HXB2 AA	HXB2 Pos	Sample AA	Sample Codon	Sample Frequency	Sample Coverage	DRM
GTC	V	3	I	ATC	0.97	6001	
AGT	S	37	N	AAT	0.9	6001	

Reverse Transcriptase

HXB2 Codon	HXB2 AA	HXB2 Pos	Sample AA	Sample Codon	Sample Frequency	Sample Coverage	DRM
ATG	M	41	L	TTG	0.0094	6001	NRTI S
AAA	K	65	R	AGA	0.0096	6001	NNRTI S + NRTI S
AAG	K	102	Q	CAG	0.95	6001	
GAA	E	122	K	AAA	0.9	6001	
AGT	S	162	C	TGT	0.96	6001	
TAT	Y	181	C	TGT	0.0096	6001	NNRTI + NRTI
GGA	G	190	A	GCA	0.0099	6001	NNRTI + NRTI
CTT	L	214	F	TTT	0.91	6001	

Reverse Transcriptase

HXB2 Codon	HXB2 AA	HXB2 Pos	Sample AA	Sample Codon	Sample Frequency	Sample Coverage	DRM
ATG	M	41	L	TTG	0.0094	6001	NRTI S
AAA	K	65	R	AGA	0.0096	6001	NNRTI S + NRTI S

Pos	A	C	G	T	-
-3	**5878**	0	0	0	123
-2	**5815**	0	6	0	180
-1	13	0	**5888**	0	100
0	**5089**	0	4	0	908
1	**5757**	0	58	0	186
2	**5841**	0	2	0	158
3	**5830**	0	0	0	171
4	**5822**	0	0	0	179
5	**5445**	0	2	0	554

AAG	K	102	Q	CAG	0.95	6001	

JULIET EXAMPLE: WHY PHASING MATTERS

unknown							
unknown			Sample Variants				
Codon	AA	Pos	AA	Codon	%	Coverage	DRM
A C G	T	**30**	M	A T G	40	1095	
C T G	L	**36**	M	A T G	39	1093	

unknown								A	B	C	D
unknown			Sample Variants					Haplotypes %			
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	54.1	34	6.1	5.9
A C G	T	**30**	M	A T G	40	1095					
C T G	L	**36**	M	A T G	39	1093					

JULIET EXAMPLE: HIV POL

Reverse Transcriptase								A	B	C	D	E	F	G	H	I
HXB2			Sample Variants					Haplotypes %								
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
ATG	M	41	L	TTG	0.99	2903	ABC + DDI + TDF + D4T + ZDV									
AAA	K	65	R	AGA	1	2577	3TC + FTC + ABC + DDI + TDF + D4T									
GGG	G	99	G	GGT	0.72	2907										
TTA	L	100	F	TTT	0.85	2819	MGI									
TAT	Y	181	C	TGT	0.95	2939	NVP + EFV + ETR + RPV									

Pos	A	C	G	T	-	N
-3	**2940**	6	0	1	4	0
-2	0	0	5	**2941**	5	0
-1	0	**2945**	0	1	5	0
0	0	0	1	**2944**	6	0
1	**2915**	4	28	1	3	0
2	0	0	2	**2946**	3	0
3	1	**2945**	0	2	3	0
4	**2905**	3	0	0	43	0
5	**2943**	5	1	2	0	0

GGA	G	190	A	GCA	1	2941	MGI + NVP + EFV + ETR + RPV									
ACC	T	215	Y	TAC	0.88	2940	ABC + DDI + TDF + D4T + ZDV									

JULIET EXAMPLE: HIV POL *IN-SILICO* MIX

Protease

HXB2			Sample Variants					A	B	C	D	E	F	G	H	I
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	Haplotypes %								
								92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
CGA	R	**8**	X	TGA	0.98	2931	MGI									

Reverse Transcriptase

HXB2			Sample Variants					A	B	C	D	E	F	G	H	I
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	Haplotypes %								
								92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
ATG	M	**41**	L	TTG	0.99	2903	ABC + DDI + TDF + D4T + ZDV									
AAA	K	**65**	R	AGA	1	2577	3TC + FTC + ABC + DDI + TDF + D4T									
GGG	G	**99**	G	GGT	0.72	2907										
TTA	L	**100**	F	TTT	0.85	2819	MGI									
TAT	Y	**181**	C	TGT	0.95	2939	NVP + EFV + ETR + RPV									
GGA	G	**190**	A	GCA	1	2941	MGI + NVP + EFV + ETR + RPV									
ACC	T	**215**	Y	TAC	0.88	2940	ABC + DDI + TDF + D4T + ZDV									

Integrase

HXB2			Sample Variants					A	B	C	D	E	F	G	H	I
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	Haplotypes %								
								92.5	1.2	1.2	1	1	0.8	0.8	0.8	0.7
AAA	K	**188**	K	AAG	0.92	2923	MGI									

JULIET EXAMPLE: PHI29 *IN-VITRO* MIX

ORF								A	B	C	D	E	F	G
PHI29			Sample Variants					Haplotypes %						
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	82.7	4	2.9	2.8	2.6	2.6	2.4
TGG	W	**99**	Q	CAG	14	25018								
			Y	TAC	2.6									
CAC	H	**116**	N	AAC	2.7	24609								
TAT	Y	**148**	Y	TAC	2.7	26368								
AAG	K	**150**	T	ACG	2.8	25410								
TAC	Y	**156**	F	TTC	2.3	25979								
			H	CAT	2.8									
AAA	K	**157**	E	GAG	3.1	23755								
GAG	E	**418**	K	AAG	3.5	26065								
GAC	D	**520**	K	AAG	2.5	26488								
GAG	E	**539**	K	AAG	4.6	25997								

JULIET EXAMPLE: HIV-1 K103N

A A A	K	**103**	N	A A C	15	25518	NVP + EFV

Pos	A	C	G	T	-	N
-3	**27263**	2	17	0	0	0
-2	**27063**	1	219	0	0	0
-1	447	0	**26149**	1	686	0
0	**25386**	3	129	0	1765	0
1	**27223**	13	31	1	15	0
2	**23167**	4102	11	2	1	0
3	**27247**	32	3	1	0	0
4	**27255**	5	20	3	0	0
5	**27233**	6	28	14	0	0

G T A	V	**106**	V	G T G	99	27256	
A C A	T	**107**	T	A C C	8.8	27217	

JULIET EXAMPLE: ONCOLOGY SCREENING

EGFR

HG38			Sample Variants					Haplotypes %			
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	A 54.2	B 33.6	C 6.1	D 6
A C G	T	30	M	A T G	39	1085					
C T G	L	36	M	A T G	39	1083					

BRAF

HG38			Sample Variants					Haplotypes %	
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	A 84.5	B 15.5
T G A	X	32	X	T A A	15	1176			

PTPN11

HG38			Sample Variants					Haplotypes %	
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	A 97	B 3
A G G	R	26	S	A G T	3	698			

JULIET EXAMPLE: BCR-ABL KINASE DOMAIN PATIENT SAMPLE

Juliet summary

▼ Input data

Input file: `/Users/atoepfer/bcr/bcr_cleric.bam`
Command line call: `./juliet /Users/atoepfer/bcr/bcr_cleric.bam --mode-phasing --config-file /Users/atoepfer/bcr/bcr_config.json`
Juliet version: `1.3.1 (commit 78ec7e3-dirty)`

▼ Variant Discovery

▶ Legend

ABL1								A	B	C	D	E	F	G	H	I	J	K	L	M
NM_005157.5			Sample Variants					Haplotypes %												
Codon	AA	Pos	AA	Codon	%	Coverage	DRM	56	17.5	4.4	4	3.9	3.8	3.6	2.5	1	1	0.9	0.8	0.7
ATG	M	244	V	GTG	5.9	2698														
GGG	G	250	E	GAG	1.4	1871														
CAG	Q	252	H	CAC	0.88	2701														
GAG	E	255	K	AAG	4.6	2619	imatinib + nilotinib													
			M	ATG	1.3															
			V	GTG	61		imatinib + nilotinib													
			X	TAG	0.95															
ACT	T	315	I	ATT	8.5	2677	imatinib + dasatinib + nilotinib + bosutinib													
GAG	E	355	G	GGG	7.6	2558														

▼ Drug Summaries

Drug	Gene	Reference AA	Reference Pos	Sample AA	Sample %
bosutinib	ABL1	T	315	I	8.5
dasatinib	ABL1	T	315	I	8.5
imatinib	ABL1	E	255	K	4.6
		E	255	V	61
		T	315	I	8.5
nilotinib	ABL1	E	255	K	4.6
		E	255	V	61
		T	315	I	8.5

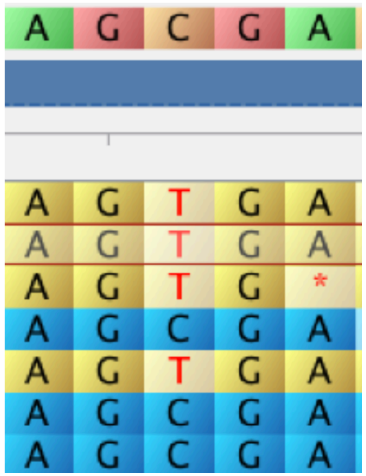
TARGET CONFIGURATION FOR BCR-ABL

```
{
    "referenceName": "NM_005157.5",
    "referenceSequence": "TTAACAGGCGCGTCCC...",
    "genes": [
        {
            "name": "ABL1",
            "begin": 193,
            "end": 3585,
            "drms": [
                {
                    "name": "imatinib",
                    "positions": [ "315A","Y253","255KV","V299L","F317AICLV" ]
                },
                {
                    "name": "dasatinib",
                    "positions": [ "T315AI","V299L","F317AICLV" ]
                }
            ]
        }
    ]
}
```

MINOR VARIANT OR PCR HETERODUPLEXES?

(1)

Subread for a CCS read with a heteroduplex



forward and reverse strand

(2)

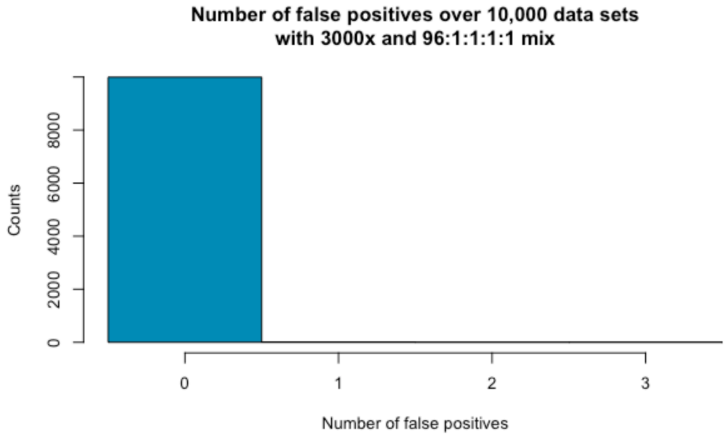
Filtering via QVs and conversion to N

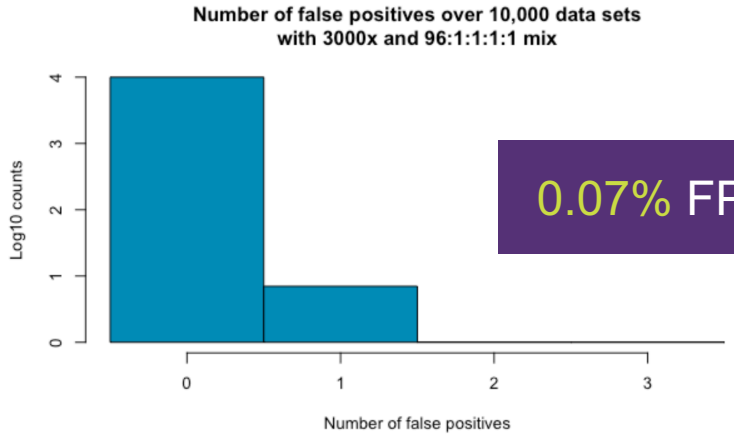
TAT	Y	181	C	TGT	0.95	2933	NVP + EFV + ETR + RPV

Pos	A	C	G	T	-	N
-3	**2933**	0	0	0	2	63
-2	0	0	0	**2975**	0	23
-1	0	**2967**	0	0	1	30
0	0	0	0	**2979**	1	18
1	**2935**	0	29	0	0	34
2	1	0	0	**2981**	0	16
3	0	**2976**	0	0	0	22
4	**2897**	0	0	0	31	70
5	**2899**	0	0	0	0	99

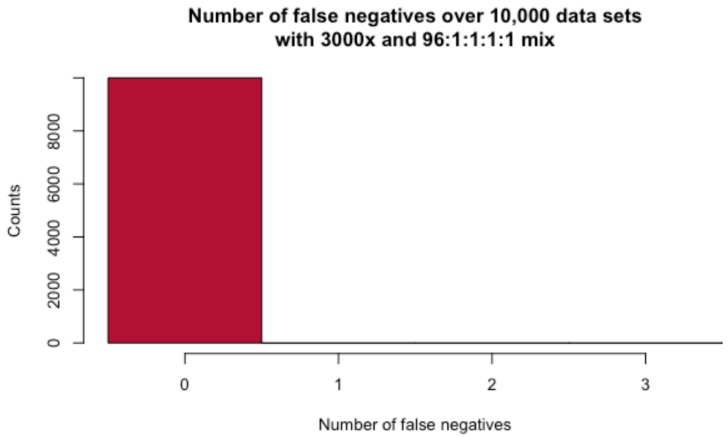
PERFORMANCE BENCHMARK

Mix *in-silico* five known HIV strains, 1-3 SNVs distance
96% WT, 1% minors
10,000 data sets at 3000x





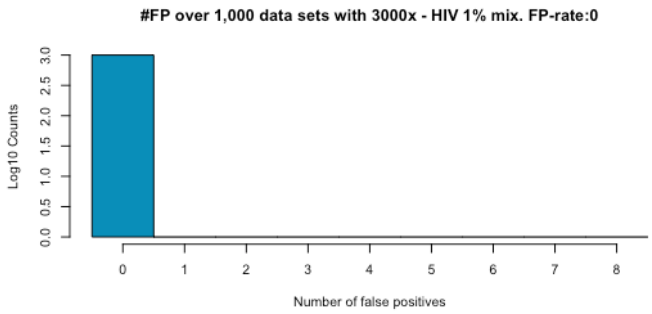
0.07% FP rate

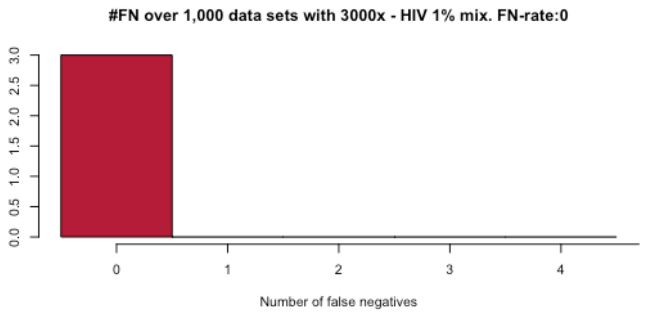


0.0% FN rate

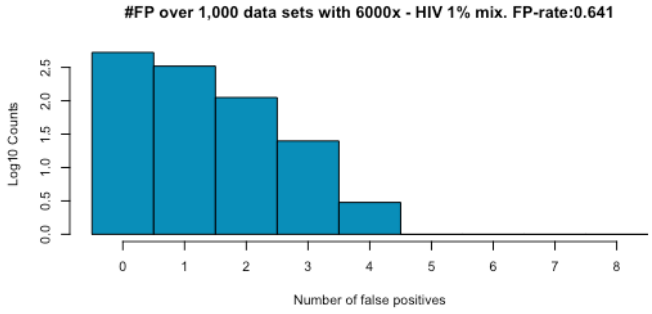
PERFORMANCE BENCHMARK – HIV – RT/PCR ERRORS

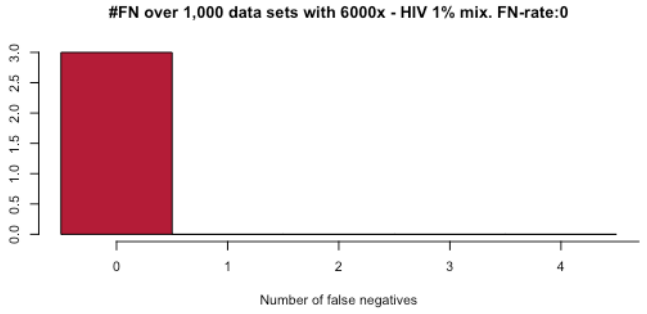
0 FPs / sample



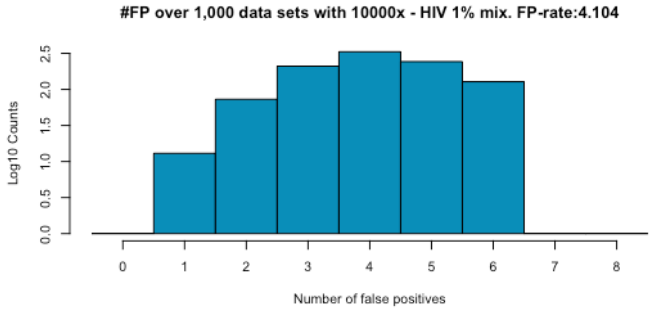


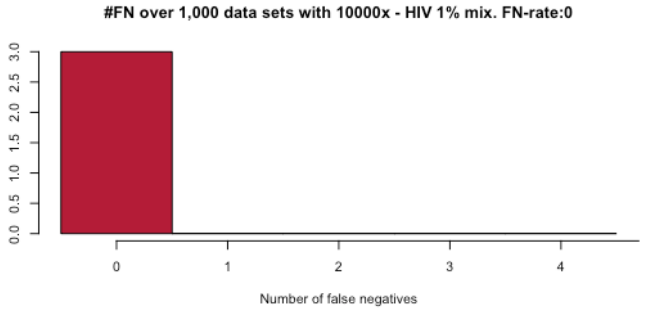
0.6 FPs / sample





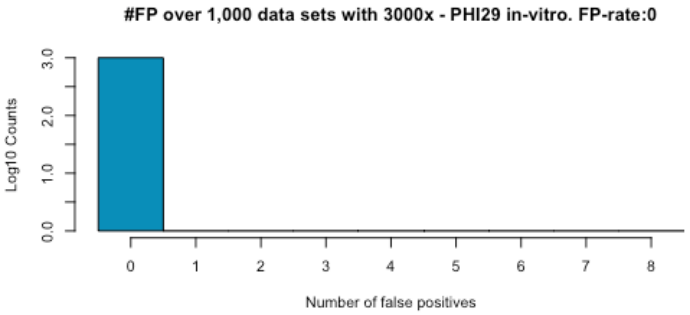
4.1 FPs / sample

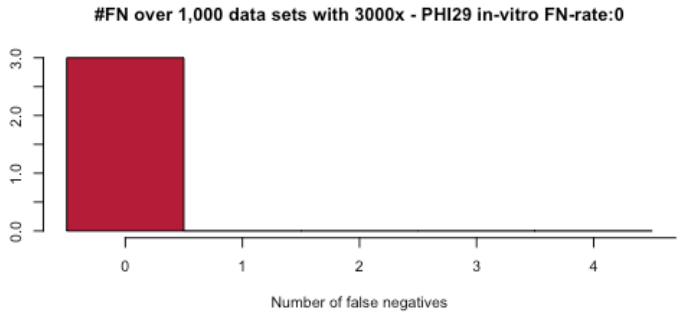




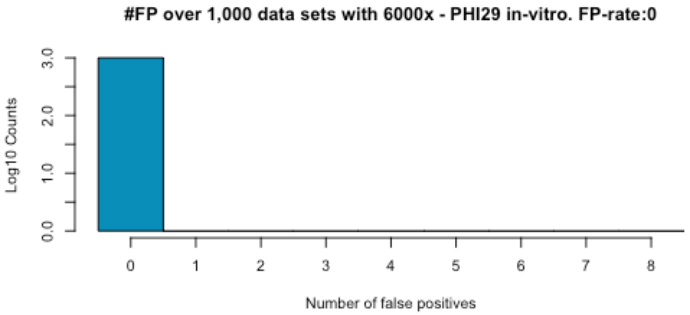
PERFORMANCE BENCHMARK – PHI29 IN-VITRO NO PCR ERRORS

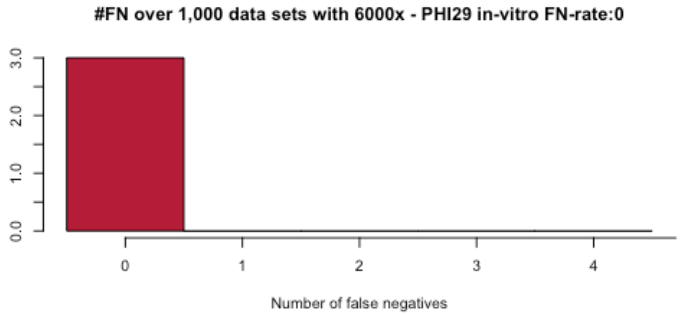
0 FPs / sample



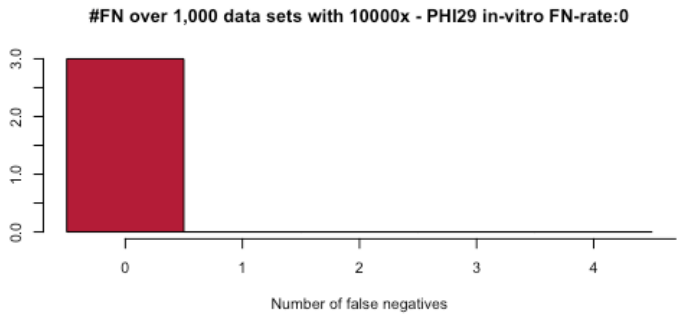


0 FPs / sample





0 FPs / sample



JULIETFLOW – ONE CLICK ANALYSIS

```
$ julietflow                    \
    -i movie.ccs.bam            \
    -r reference.fasta          \
    -c target_config.json
```

AVAILABILITY:

- Official release @SMRT Link 5.0 (soon)
- “as-is” version @github.com/PacificBiosciences/minorseq

Documentation

- Getting Started
- Projects
  - Minor Variant Calling `juliet`
  - Reduce Alignment `fuse`
  - Swap Alignment Reference `cleric`
  - Minor Variant Pipeline `julietflow`
- Running your sample 101
- Developer environment
- PacBio open source license

Quick Tools Overview

Minor variant caller

`juliet` identifies minor variants from aligned ccs reads.

Reduce alignment

`fuse` reduces an alignment into its closest representative sequence.

Swap BAM reference

`cleric` swaps the reference of an alignment by transitive alignment.

Minor variant pipeline

`julietflow` automatizes the minor variant pipeline.

Disclaimer

The GitHub version is an "as-is" develop distribution of minorseq and not the official release of the software; a coherent git history cannot be guaranteed.

Help

Official support is only provided for official and stable SMRT Analysis builds provided by PacBio and not for source builds. Unofficial support for source builds is provided via github issues.

PACBIO®

www.pacb.com

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2017 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies.

All other trademarks are the sole property of their respective owners.