

UCSF

UC San Francisco Previously Published Works

Title

MinorityReport, software for generalized analysis of causal genetic variants

Permalink

<https://escholarship.org/uc/item/8h71m4wv>

Journal

Malaria Journal, 16(1)

ISSN

1475-2875

Authors

Horst, Jeremy A

Wu, Wesley

DeRisi, Joseph L

Publication Date

2017-12-01

DOI

10.1186/s12936-017-1730-2

Peer reviewed

METHODOLOGY

Open Access



MinorityReport, software for generalized analysis of causal genetic variants

Jeremy A. Horst^{1*}, Wesley Wu¹ and Joseph L. DeRisi^{1,2}

Abstract

Background: The widespread availability of next generation genome sequencing technologies has enabled a wide range of variant detection applications, especially in cancer and inborn genetic disorders. For model systems and microorganisms, the same technology may be used to discover the causative mutations for any phenotype, including those generated in response to chemical perturbation. In the case of pathogenic organisms, these approaches have allowed the determination of drug targets by means of resistance selection followed by genome sequencing.

Results: MinorityReport is open source software written in python that facilitates the comparison of any two sets of genome alignments for the purpose of rapidly identifying the spectrum of nonsynonymous changes, insertions or deletions, and copy number variations in a presumed mutant relative to its parent. Specifically, MinorityReport relates mapped sequence reads in SAM format output from any alignment tool for both the mutant and parent genome, relative to a reference genome, and produces the set of variants that distinguishes the mutant from the parent, all presented in an intuitive, straightforward report format. MinorityReport features tunable parameters for evaluating evidence and a scoring system that prioritizes reported variants based on relative proportions of read counts supporting the variant in the mutant versus parent data sets. The utility of MinorityReport is demonstrated using previously published publicly available data sets to find the determinants of resistance for novel anti-malarial drugs.

Conclusions: MinorityReport is readily available (github: JeremyHorst/MinorityReport) to identify the genetic mechanisms of drug resistance in *Plasmodium*, genotype-phenotype relationships in human diads, or genomic variations between any two related organisms.

Keywords: Bioinformatics, Mutation, Genotype-phenotype, Drug mechanism, Genetic etiology, CNV, Missense

Background

Sequencing to characterize genotype-phenotype relationships

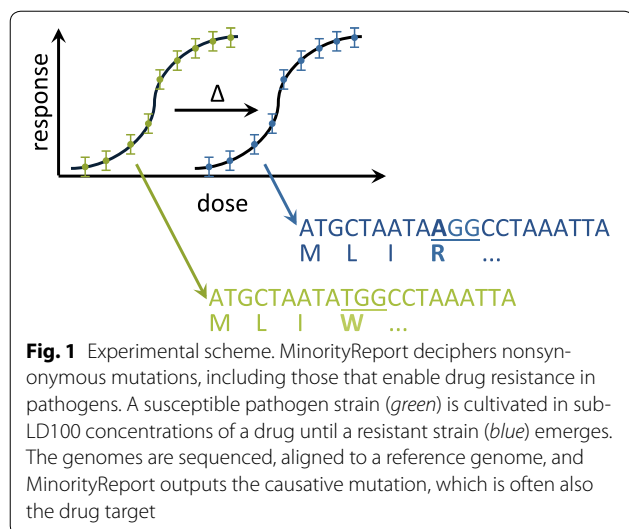
The advent of robust high throughput genetic sequencing methods has made genome sequencing tractable for any sample from any organism. Comparison of genomes for related organisms (within the same lineage of a species) with different physiologic, behavioural, or anatomic phenotypes has been increasingly used to characterize the genetic determinants of the feature, i.e. genotype-phenotype relationships. Differences in the genome sequences

of two related organisms with a different phenotype describe the possible causal genetic variants (see Fig. 1). This technique has been used to find the molecular determinants of resistance for novel compounds against the causative agent of the most deadly form of malaria, *Plasmodium falciparum* [1], including a drug now in clinical trials [2], and an inhibitor of the apicoplast—an organelle unique to Apicomplexa [3]. The same principles have been used to find novel mutant genes underlying human craniofacial congenital malformations including craniosynostosis [4], auriculocondylar syndrome and thereby the universal signal for lower jaw patterning [5], and acromelic frontonasal dysostosis [6]. Beyond congenital anomalies and drug targets, this experimental system is relevant to any genotype-phenotype study.

*Correspondence: jahorst@gmail.com

¹ Department of Biochemistry and Biophysics, University of California San Francisco School of Medicine, 1700 4th St, QB3 Room 404, San Francisco, CA 94158-2330, USA

Full list of author information is available at the end of the article



How to

To perform this analysis, genomic sequencing reads from each organism (e.g. in FASTQ format) are first aligned to the reference genome FASTA file, resulting in Sequence Alignment/Map (SAM) format files [7]. The observed variations from the reference genome are then evaluated for sufficient supporting evidence by quantity and proportion of reads in the mutant, and support for non-variation in the parent. The variants that pass these threshold filters are then placed in the context of coding sequences, defined in gene file format files (GFF3), to determine whether the variant changes the gene product, i.e. is nonsynonymous (Fig. 2). Variants observed consistently in multiple independent pairs of organisms that differ in the same manner are likely to be the causal genetic variant that underlies the phenotype. In the case of drug resistance selections, the underlying assumption for a mutation or copy-number variant which leads to resistance is that it will be under purifying selection unlike other variants present in the population at the time of the initial selection.

This process requires the reference genome and gene model, and deep sequencing data for unaffected parents and the affected child for humans, or similarly the drug-sensitive parent strain and the derivative drug-resistant mutants for pathogens. The reference genome is required as a contextual template to align the sequencing reads and thereby identify genetic variants. The gene model is needed to identify whether a variant is in a protein-coding region, and whether the variant changes the resulting protein sequence.

While this process is straightforward in concept, many bioinformatic analysis computer programs work only on one or an arbitrary set of organisms, e.g. human and

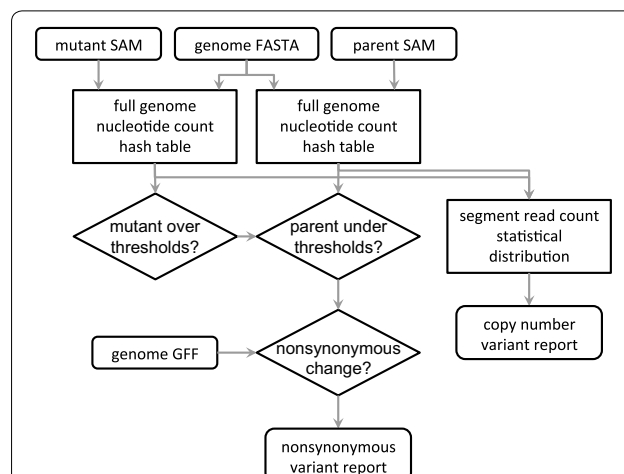


Fig. 2 Software flow chart. MinorityReport relates mapped sequence reads in SAM format output from any alignment tool for both the mutant and parent genome, relative to a reference genome, to the protein coding sequence, and produces the set of nonsynonymous variant that distinguish the mutant from the parent

mouse, or do not integrate both mutation and copy-number variant analysis into a unified pipeline. A few robust generalized tools are available to call variants between one set of sequencing reads against any reference genome, and a subset of these call nonsynonymous variants. However, these tools require extensive data preparation, have software library dependencies that complicate installation, are sensitive to the many imperfections in gene models for the many poorly characterized organisms (i.e. they fail to complete running or give any output), and they do not perform comparative analysis between multiple sequencing sets. In practice, many researchers cobble together pipelines based on several different tools to accomplish this task.

Freedom of alignment

Alignment algorithms continue to evolve at rapid pace, and the most recent iterations capture alignments to much more dissimilar sequences with inexact matches and significant insertions or deletions (indels). e.g. Bow-Tie2, handle indels of limited size, and further sequence differences [8]. Algorithms such as STAR [9] and GSNAP [10] handle extensive indels, and highly distinct but still recognizably similar read-genome sequence matches. These advances are useful to address the genotype-phenotype problem. Fortunately, these and other genome alignment algorithms have converged on a standard SAM format [7] which enable downstream applications, such as the one described here, to be independent of the base aligner.

Copy number variants

Another major type of genetic anomaly that can underlie genotype-phenotype relationships is a gene copy number variant (CNV). When the function of a gene becomes particularly adaptive, useless, or harmful, environmental pressures can select for the replication or loss of any number of copies of a gene or set of genes [11]. Because CNVs affect gene dosage, they have the potential for physiologic impact. Various approaches are used to detect CNVs on a large scale, with the primary clinical tool being array comparative genomic hybridization [12]. Algorithms have been constructed to use indirect evidence to evaluate CNVs, such as inference through analysis of single nucleotide polymorphism (SNP) chips [13]. However, as with all microarray technologies, next generation genome sequencing technologies are replacing the use of array CGH and SNP chips due to their portability to any biological system, and unbiased sequence data generation. The normalized abundance of sequencing reads that map to each gene is a direct measure of the number of copies of the gene. CNVs can be estimated using sparse read coverage data sets as well [11]. However, the enormous number of reads accessible through contemporary sequencing instruments (e.g. the Illumina HiSeq series) makes direct assessment through deep coverage straightforward.

Results and discussion

MinorityReport facilitates the comparison of any two sets of genome alignments for the purpose of rapidly identifying the spectrum of nonsynonymous or indel changes in a presumed mutant relative to its parent. In addition, the data model for detecting nonsynonymous variants integrates rapid assessment of CNVs. Tunable parameters for both analyses enable reporting of high or low purity nonsynonymous mutations, and large or small range CNV detection. This software does not address noncoding variants such as promoter or enhancer element mutations, intergenic variants that may occur in encoded functional RNAs, nor chromosomal rearrangements, which may be assessed with the same sequencing data. Algorithms deployed in this software are described, and utility is demonstrated on freely available data sets that carry the genetic determinants of resistance for anti-malarial drugs.

Nonsynonymous variant caller algorithm

A simplified flow chart of the nonsynonymous variant caller algorithm in MinorityReport is shown in Fig. 2. FASTA and GFF3 files for the reference genome are taken as input to build the gene model. Indices for genes, exons,

and splice variants are taken from coding sequence lines (those with CDS entered in the phase column). Gene descriptions are taken from gene lines. Reverse complement indices are translated for genes present on the negative strand (python list indices are inclusive at the beginning but exclusive at the end, so must be shifted 1 nucleotide towards the 3' direction). Sequences are found from these indices in the reference genome FASTA file, and checked for start and stop codons. Next, a hash table (python dictionary) is made from the reference genome FASTA file to store the sequence evidence: for each sequence position in each chromosome a key is created for each of the possible 4 nucleotides, and a zero value is entered (indels are added later).

Sequence Alignment/Map files of sequencing reads aligned to the reference genome for the parent and mutant are each taken as input and applied into separate sequence evidence hash tables. For each mapped read pair, the mate pair read is checked for matched mapping location. The CIGAR entry is read for presence of insertions or deletions (indels) in the read. The mapped strand is extracted from the binary flag entry. Finally, each nucleotide in each paired read is added into the sequence evidence hash table. Insertions and deletions are added to the position entry as they occur, and counted the same way as nucleotides. An example position: "sequence_evidence['chr7'] [1423712] = {'A':3, 'C':5, 'G':51, 'T':3, '-':426, 'AACTAC':20}" where chr7 is the chromosome, 1423712 is the position on chromosome 7, and 3 A's, 5 C's, 51 G's, 3 T's, 426 deletions, and 20 AACTAC insertions are observed at the position. Insertions are denoted in the direction of the positive strand; the first nucleotide in the insertion entry is technically not part of the insertion, but is counted in the insertion to enable a consistent notation system.

Next, the sequence evidence hash tables are scanned for nonsynonymous variants that appear in the mutant but not the parent. Chromosomes are read in turn. Then evidence thresholds are applied to each entry of the position in a progression to maximize efficiency: variant of the reference genome, supra-threshold number of mutant reads and fraction of reads at the position support the variant, supra-threshold number of parent read coverage, subthreshold number of parent reads and fraction of reads at the position support the variant, the variant is in a coding sequence region, and the variant changes an amino acid in the translated protein sequence (nonsynonymous). Variants passing these criteria are reported (Table 1). Each reported variant is assigned a priority score as follows:

$$priority = \frac{\left(\frac{mutation_{mutant} + position_{mutant}}{\mu_{mutant}} \right) + \left(\frac{position_{parent}}{\mu_{parent}} \right) - \left(\frac{nonmutation_{mutant}}{position_{mutant}} \right)}{1 + \left(\frac{mutation_{parent}}{\mu_{parent}} \right)}$$

Table 1 Sample MinorityReport output

Sample	Score	Chromosome	Genomic mutation	Gene mutation	Protein mutation	Mutant		Parent		Gene id	Strand	Protein length	Gene description		
						Mutation counts	Mutation proportion	Mutation counts	Mutation proportion						
S71	4.7	chr12	G531532A	C1267T	A423V	43	0.977	44	12	0.017	693	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
	3.8	chr12	G531566T	C1233A	P412T	30	1	30	3	0.004	684	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
	3.3	chr6	C1111087A	C2890A	A964D	31	1	31	3	0.006	522	PF3D7_0627700	+	1147	Transportin
S72	5.5	chr12	A531751T	T1048A	L350H	89	0.989	90	2	0.003	729	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
S81	3.9	chr12	A531556T	T1243A	V415D	34	1	34	3	0.005	654	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
S82	6.4	chr12	A531751T	T1048A	L350H	52	0.981	53	2	0.003	729	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
S83	4.3	chr12	G531566T	C1233A	P412T	72	0.973	74	3	0.004	684	PF3D7_1211900	–	1264	Non-SERCA-type Ca2 Transporting P-ATPase (ATP4)
	3.2	chr6	C1111087A	C2890A	A964D	60	1	60	3	0.006	522	PF3D7_0627700	+	1147	Transportin

MinorityReport output can be combined across multiple strains (s71, s72, s81, s82, s83) derived from the same parent strain in independent resistance experiments. The priority score aids identification of variants that reach near purity in mutant strains relative to the parent strain, with strong supporting evidence. The gene containing a mutation in all strains, here ATPase4, is the known determinant of resistance. Shown are all variants reported when using the default thresholds

$mutation_{mutant}$ refers to the quantity of sequencing reads that contain the variant in the mutant sample, and similarly $mutation_{parent}$ refers to the same in the parent sample. $nonmutation_{mutant}$ refers to the quantity of reads that contain anything other than the variant in the mutant sample. $position$ variables refer to the total number of reads overlapping the position of the variant in each sample, and μ refers to the mean number of reads overlapping any position in the genome for each sample.

The rationale for the terms in the priority score calculation is as follows. The relative abundance of sequencing reads that identify a nucleotide or indel (variant) comprises the evidence for the variant. Reads that identify any other allele at the position, and reads identifying the same variant in the parent strain comprise the evidence against the variant being novel and causal. Low coverage in either the parent or mutant samples would decrease confidence in the variant, so high relative coverage of the samples increases prioritization.

CNV caller algorithm

As each position in each chromosome is scanned for nonsynonymous variants unique to the mutant, the total counts in each of parent and mutant for each position are applied to a CNV evidence model. After the nonsynonymous variant scan is complete for each chromosome, the genome is evaluated for regions with significant differences in median read coverage between mutant and parent. The probability of the parent or mutant being significantly high or low in copy number (gene abundance) is estimated by relating the number of reads that map to the position to the mean and standard deviation for the chromosome, i.e. the z-score. Positions are considered significant when a number of reads are mapped in the mutant that correspond to being above or below a probability threshold or number of copies, with respect to the parent.

Copy number variants are calculated as the ratio of reads that map to each sliding window in the mutant and parent data sets, normalized by the total number of reads for the parent and mutant. The probability of observing a CNV at random for each variant is estimated by placing the log2 transform in the context of log2 normalized read ratios for all sliding windows assessed across the genome. The resulting z-score ($z = (x - \text{mean}) / \text{standard deviation}$) is transformed to a probability estimate by the Gaussian transformation.

The CNV-seq algorithm is widely used to calculate CNVs from sequencing data [11]. The CNV-seq algorithm calculates the probability for each observed CNV occurring at random using the covariation of the read numbers in the parent and mutant data sets, with the Geary-Hinkley transformation. As seen in Figs. 3, 4 and

5, read coverage is not homogenous. Areas with deviations in coverage due to external processing factors, such as the non-uniform distribution of so-called “random” hexamer-nucleotide primers, or the inaccuracy of genome alignments, may become inflated or diminished using this approach. The total amounts of reads in each set are also the result of processing factors. In attempt to simplify, the statistical probability of observing read ratios normalized by the total read count is calculated by applying a Gaussian distribution to the entire read set, i.e. including all chromosomes. Recent work analysing genetic crosses of *P. falciparum* laboratory strains employed a hidden Markov permutation of the Gaussian distribution to improve CNV significance estimation [14]. While read coverage does not follow a normal distribution, at present, hidden Markov modelling requires library dependencies that are obtrusive to use, and so the software presently uses a less ideal distribution model.

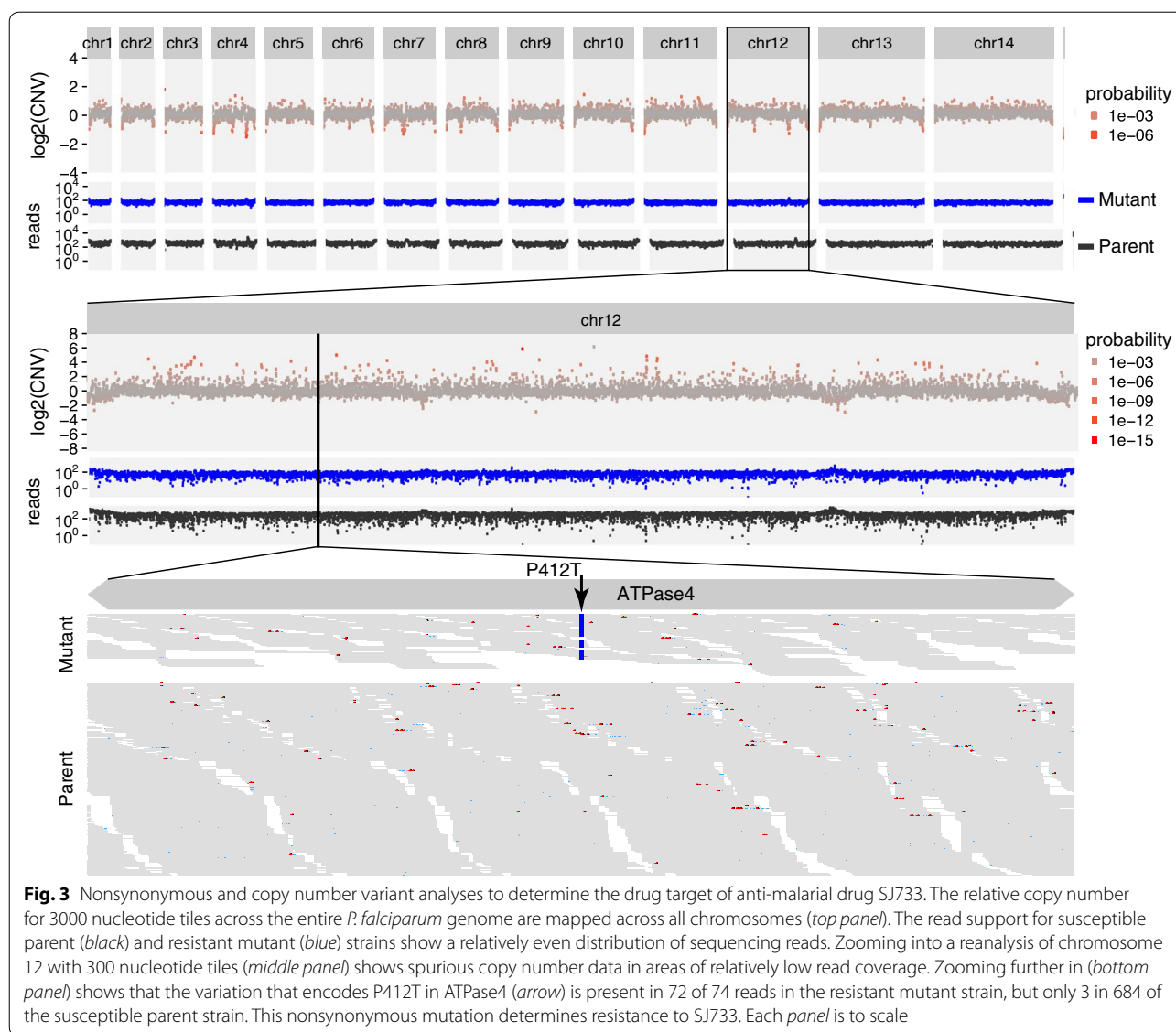
Application to determine target of novel anti-malarial compounds

Compound SJ733

The discovery of a novel anti-malarial agent was recently reported [2], and is now being tested for safety in humans. A key milestone for understanding the mechanism of action for SJ733 and its derivatives was to identify molecular determinants of resistance. Resistant sub-clones of a susceptible parent strain were selected by long incubations in concentrations of the compound that allowed very few organisms to live, and identified the resistance loci by comparing genome sequence data between parent and resistant sub-clones (mutants). Comparing genomic sequencing reads between parent and mutants revealed a few variants in most mutants, while only one gene contained variants in all mutants: the gene encoding PfATP4, a sodium-dependent ATPase transporter [2].

The process of simultaneously assessing copy number and nonsynonymous variation in the search for the mutation that enables resistance to SJ733 is illustrated in Fig. 3. For a haploid genome under drug selection, the underlying genetic assumption is that the causal mutant allele will be driven to purity in the population of surviving cells, and thus it is reasonable to set an absolute threshold of at least 90% for the proportion of reads that are variant at any given position. This threshold is a tunable parameter and would clearly need to be below 50% for a diploid genome, for example.

The sequencing data for this experiment are available through the NCBI Short Read Archive (SRA): SRX644366 is the mutant S83, SRX643297 is the parent strain, and other resistant strains derived in parallel independent experiments can be found through SRP043648.

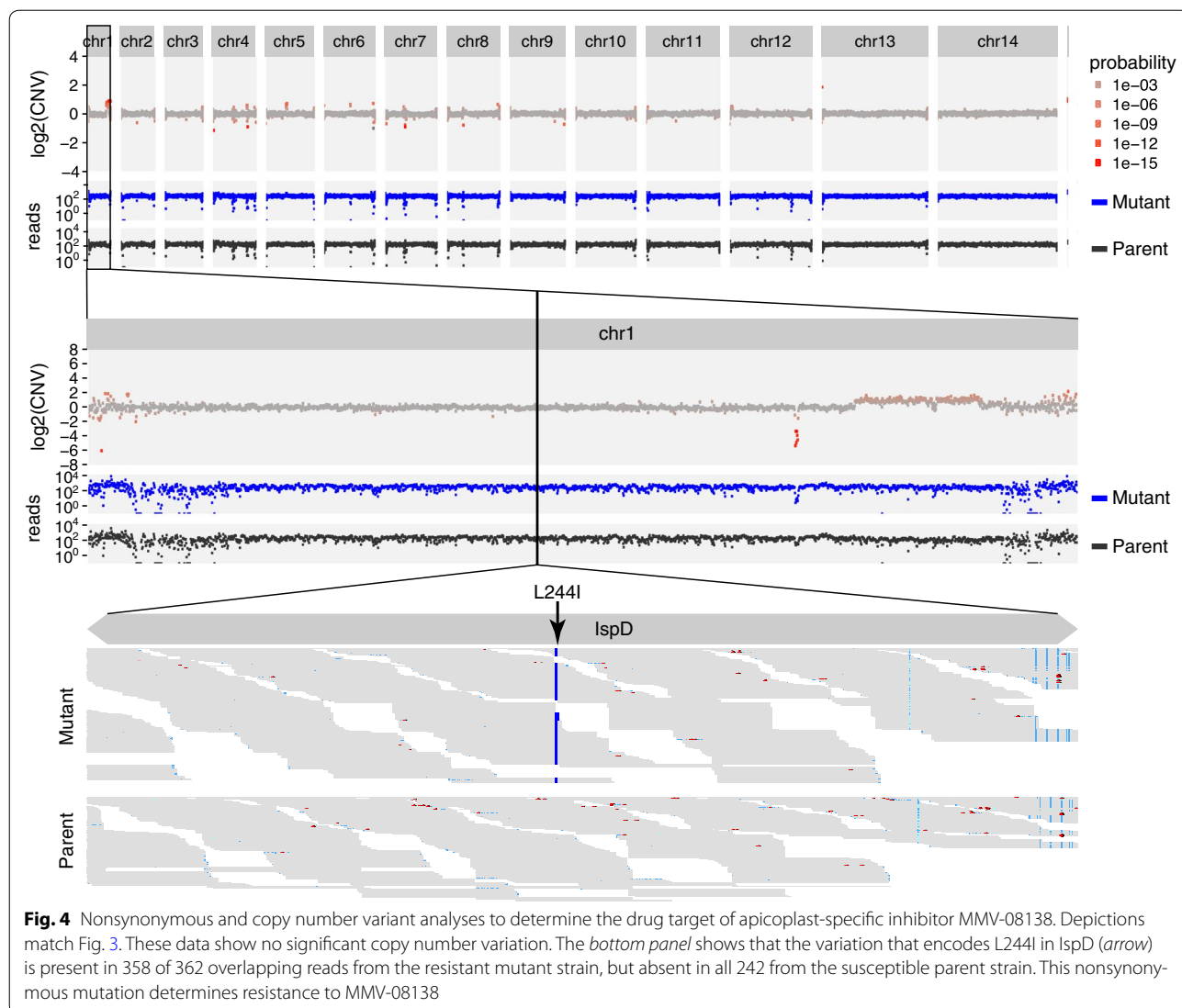


The list of nonsynonymous variants found by MinorityReport (Table 1) demonstrates the coincident emergence of mutations in the same gene from 5 independent selection experiments. In fact, in 3 of 5 selections, the only mutation that passes these basic thresholds are those in PfATP4 [2]. Note that relaxing the mutant proportion threshold would reveal several high scoring nonsynonymous variants, but essentially all of these belong to highly repeated gene families, such as the PfEMP1 family of antigenic variation genes, which are inherently prone to spurious alignment. The appearance of these non-causal variants, such as transportin, in the data set exemplifies the rare occurrence of real or artifact passenger mutations that may be reported, and also emphasizes the utility of analysing several independent selections.

Also, the CNV assessments differ between the granular genome-wide scan performed with 3000 nucleotide-long tiles (top) versus the chromosome-specific scan performed with 300 nucleotide-long tiles (middle); while this was done to construct an illustrative figure with a reasonable amount of data, it shows that tile size affects the appearance of copy number and likelihood calculations: larger tiles are recommended for more relevant probability values. In the case of PfATP4, our analysis produced no evidence of copy number variation associated with resistance to the compound.

Apicoplast inhibitor MMV-08138

The mechanism of action of a *Plasmodium* apicoplast inhibitor MMV-08138 was recently reported [3]. As with



SJ733, the genome sequence of sub-clones raised for resistance to the drug were compared against the susceptible parent strain. A single nonsynonymous variation (IspD L244I) not observed in the parent strain emerged, and reached nearly complete (98.9%) purity in the resistant mutant shown in Fig. 4. With the default threshold settings, 23 putative variants are output for this sample by MinorityReport (Additional file 1: Table S1). As well, the causal mutant receives over twice the priority score of any other variant.

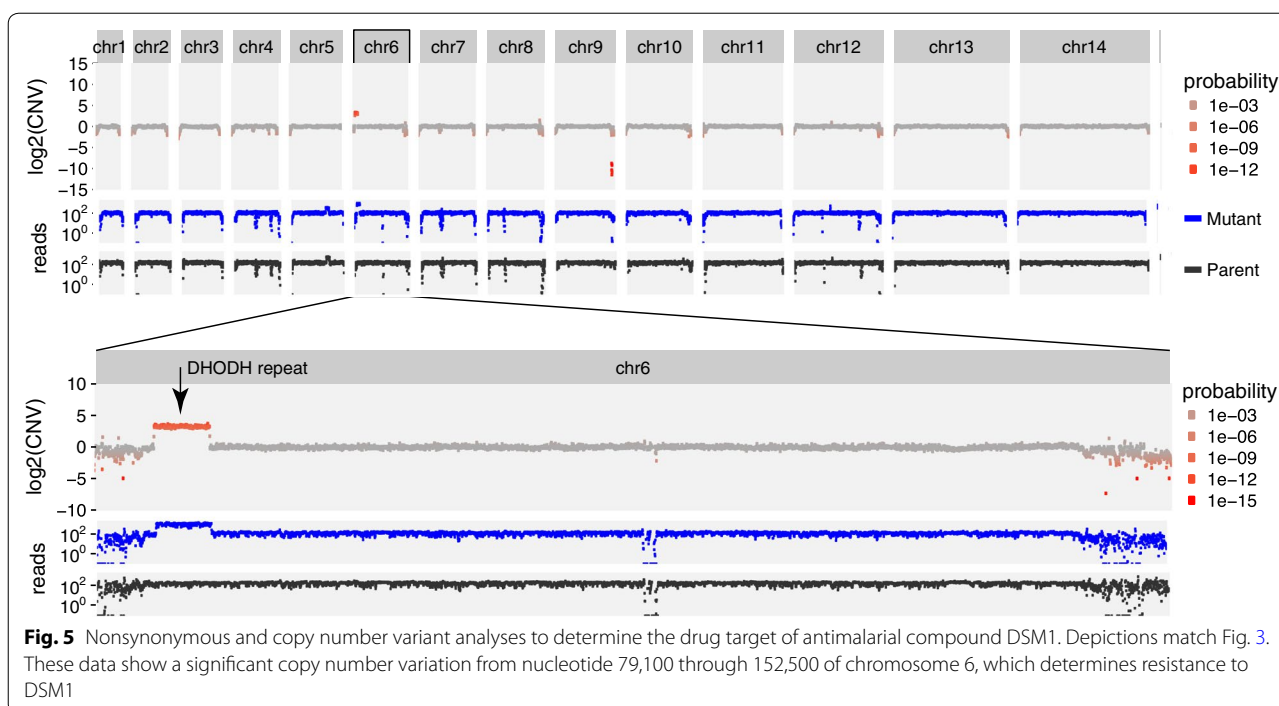
Dihydroorotate dehydrogenase inhibitor DSM1

We also previously demonstrated that *P. falciparum* can achieve drug resistance to the dihydroorotate dehydrogenase inhibitor, DSM1, and perhaps generally, by randomly amplifying a large AT-repeat flanked region of the genome, then specifically further amplifying the region

in progeny that successfully evade the drug effects [1]. Amplified regions likely contain the drug target, which can also contain nonsynonymous mutations that further aid resistance. MinorityReport successfully identified the amplified region (Fig. 5), and 4 nonsynonymous mutations (Additional file 2: Table S2) that were previously reported.

Conclusions

MinorityReport is a robust tool to rapidly identify the genetic differences between any two closely related organisms. It relates mapped sequence reads in SAM format from any alignment tool for both the mutant and parent genome, relative to a reference genome, and produces the set of variants that distinguishes the mutant from the parent. Output includes the spectrum of nonsynonymous changes, insertions or deletions, and copy



number variations in a presumed mutant relative to its parent. Tunable evidence filters prioritize reported variants based on relative proportions of read counts supporting the genetic variant in the mutant versus parent data sets. Since MinorityReport is organism agnostic, it will be useful for both haploid and diploid genomes, as well as microorganisms and viruses.

The software is freely available, open source, and only requires Python with libraries included in standard distributions (re, math, sys, and os). The R code for producing the data graphics shown in the Figures is also included in the github distribution (this has software library dependencies), and a “master” script that breaks the calculation into smaller parts by chromosome and launches them in parallel for faster calculation of nonsynonymous variants on machines with more processor cores than there are chromosomes in the genome.

Additional files

Additional file 1: Table S1. Nonsynonymous variant data for a *Plasmodium falciparum* mutant strain raised for resistance to an apicoplast inhibitor, versus the susceptible parent strain. The causal mutation, IspD L244I, reaches near purity in the mutant, is absent in the parent, and receives the highest priority score.

Additional file 2: Table S2. Nonsynonymous variant data for a *Plasmodium falciparum* mutant strain raised for resistance to a dihydroorotate dehydrogenase inhibitor, versus the susceptible parent strain. Multiple mutations arise for all resistant strains, following the copy number variation and amplification shown in Fig. 5.

Abbreviations

SAM: Sequence Alignment/Map; GFF3: gene file format files; CNV: copy-number variant; SNP: single nucleotide polymorphism; PfATP4: sodium-dependent ATPase transporter; SRA: NCBI Short Read Archive; DSM1: dihydroorotate dehydrogenase inhibitor.

Authors' contributions

JAH and JLD conceived of and designed the software and example applications. JAH wrote the software, prepared the figures, and drafted the manuscript. JLD and WW edited the manuscript and figures, tested and suggested improvements to the software. All authors read and approved the final manuscript.

Author details

¹ Department of Biochemistry and Biophysics, University of California San Francisco School of Medicine, 1700 4th St, QB3 Room 404, San Francisco, CA 94158-2330, USA. ² Chan-Zuckerberg BioHub, 499 Illinois St, San Francisco, CA 94158-2330, USA.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data are available on the NCBI Short Read Archive, and software is available as a supplement and will be deposited in GitHub upon publication. Data: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA253899/>.

Funding

JAH was funded by NIH/NIDCR training grant T32-DE007306. WW was funded by the Philippine California Advanced Research Institutes project IHITM 63. JLD was funded by the Howard Hughes Medical Institute.

Received: 23 July 2016 Accepted: 9 February 2017

Published online: 23 February 2017

References

- Guler JL, Freeman DL, Ahyong V, Patrapuvich R, White J, Gujjar R, et al. Asexual populations of the human malaria parasite, *Plasmodium falciparum*, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications. *PLoS Pathog*. 2013;9:e1003375.
- Jiménez-Díaz MB, Ebert D, Salinas Y, Pradhan A, Lehane AM, Myrand-Lapierre M-E, et al. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of *Plasmodium*. *Proc Natl Acad Sci USA*. 2014;111:E5455–62.
- Wu W, Herrera Z, Ebert D, Baska K, Cho SH, DeRisi JL, et al. A chemical rescue screen identifies a *Plasmodium falciparum* apicoplast inhibitor targeting MEP isoprenoid precursor biosynthesis. *Antimicrob Agents Chemother*. 2015;59:356–64.
- Cunningham ML, Horst JA, Rieder MJ, Hing AV, Stanaway IB, Park SS, et al. IGF1R variants associated with isolated single suture craniosynostosis. *Am J Med Genet A*. 2011;155A:91–7.
- Rieder MJ, Green GE, Park SS, Stamper BD, Gordon CT, Johnson JM, et al. A human homeotic transformation resulting from mutations in PLCB4 and GNAI3 causes auriculocondylar syndrome. *Am J Hum Genet*. 2012;90:907–14.
- Smith JD, Hing AV, Clarke CM, Johnson NM, Perez FA, Park SS, et al. Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontonasal dysostosis. *Am J Hum Genet*. 2014;95:235–40.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie2. *Nat Methods*. 2012;9:357–9.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
- Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010;26:873–81.
- Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinform*. 2009;10:80.
- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*. 1997;20:399–407.
- Tsuang DW, Millard SP, Ely B, Chi P, Wang K, Raskind WH, et al. The effect of algorithms on copy number variant detection. *PLoS ONE*. 2010;5:e14456.
- Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res*. 2016;26:1288–99.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

