# Conbase: a software for discovery of clonal somatic mutations in single cells through read phasing

Joanna Hård[1,5,*], Ezeddin Al Hakim[1,5], Marie Kindblom[1,5], Åsa Björklund[2], Ilke Demirci[1], Marta Paterlini[1], Bengt Sennblad[2], Erik Borgström[3], Patrik L Ståhl[1], Jakob Michaelsson[4], Jeff E. Mold[1,6] & Jonas Frisén[1,6,*]

[1] Department of Cell and Molecular Biology, Karolinska Institutet, Berzelius väg 35, SE-17177, Solna, Sweden

[2] Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Husargatan 3, SE-752 37 Uppsala, Sweden

[3] Scilifelab, Division of Gene Technology, KTH Royal Institute of Technology, SE-171 21, Solna, Sweden

[4] Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, SE-141 86, Stockholm, Sweden

[5] These authors contributed equally to the present work.

[6] These authors contributed equally to the present work.

[*] Corresponding author

Somatic variant analysis in single cells is challenging due to the large fraction of false positive and false negative variant calls. To overcome these obstacles we developed Conbase, a software for calling clonal somatic mutations in genome sequencing data from single cells. Conbase takes advantage of read phasing and observations across all single cells in a dataset and an unamplified bulk sample, to call clonal somatic mutations with high confidence and exclude false variant calls. This strategy mitigates effects of amplification errors as well as

other aspects of bioinformatic analysis, including alignment artefacts and an incomplete reference genome with respect to the genome of the donor and enables determination of presence or absence of mutations despite low read depth and high rates of allelic dropout (ADO) (>90%). We demonstrate the potential of Conbase by unambiguously defining the clonal relationships of single cells from two in vivo expanded T cell clones isolated directly from the peripheral blood of a healthy human donor.

Advances in DNA sequencing technologies have vastly improved our understanding of genetic variation between individuals. Deciphering variation within a given individual has been proven more difficult, as it requires the accurate identification of mutations that may exist in only a small fraction of the cells from which DNA was isolated. As such, identification of DNA mutations arising during normal somatic cell divisions has been limited to mutations that likely arose during early embryogenesis, and thus are found in a high proportion of cells distributed throughout the body[1,2]. Single cell DNA sequencing has recently been embraced as a method for understanding the degree of genetic variability that exists within an individual healthy human, and holds the promise of enabling lineage tracing in human subjects. Unfortunately, substantial technical limitations exist for analyzing sequencing libraries derived from single cells subjected to whole genome amplification (WGA), where locus dropout resulting from failed amplification limits the fraction of the genome that can be analyzed. In successfully amplified genomic regions, ADO and amplification bias result in false negative variant calls and amplification errors result in false positive variant calls[3,4,5,6]. In addition, other aspects of bioinformatic analysis result in erroneous genotype calls, including low complexity genomic sequences, failed realignment and an incomplete reference genome with respect to the genome of the donor[7].

We developed Conbase to identify clonal somatic single nucleotide variants (sSNVs) with the aim of deciphering cell lineages in human beings. The name Conbase is derived from the fact that genotyping is performed by confirming the allelic origin of bases at an observed variant site, through phasing of read files from single cells using heterozygous germline single nucleotide variants (gSNVs) identified in an unamplified bulk sample (**Fig. 1 a-c, Supplementary Fig. 1, Online Methods**). Phasing allows for determination of maternal or paternal origin of a variant, and thereby identification of which of the two alleles is informative for genotype predictions, representing the informative allele (**Fig. 1b**). To define phylogenetic relationships at the single cell level, it is important to confirm the absence of mutations in unmutated samples. False negative variant calls resulting from ADO, are eliminated by requiring for a sample to display reads originating from the informative allele in ordered to be assigned a genotype (**Fig. 1b, Supplementary Fig. 2**). The genotypes of samples that do not display reads originating from the informative allele are unknown (**Fig. 1b**). In sites where the presence of a sSNV was successfully predicted, >1 read originating from the informative allele is sufficient to infer the genotypes in samples exhibiting low read depth (**Supplementary Fig. 2**). False variants result from alignment artefacts, errors in the reference genome, copy number differences between the reference genome and the study subject, as well as amplification errors. False variants are to large extent excluded by analyzing multiple observed data during genotyping, including monitoring of the number of observed haplotypes within individual single cells in putative sSNV loci, as well as maximal expected read depth in an unamplified bulk sample and the density of mismatches against the reference genome in the region in which variants are present, as these observations are indicative of errors in the reference genome or structural variation resulting in mapping artefacts and read pile up (**Fig 1c, Supplementary Fig. 2, Online Methods**). Conbase outputs an interactive graphical html file, in which genotypes in sSNV loci are color coded based on

observations associated with confidence in predicted genotypes, including phased and total read depth support and presence or absence of ADO. Genotype statistics are summarized in a complementary tsv-file.

To evaluate Conbase, we designed an experiment using clonally expanded $CD8^+$ T cells isolated directly from a healthy human donor vaccinated with the yellow fever virus vaccine (YFV-17D). Genomic DNA samples from bulk peripheral blood cells were prepared in parallel and two separate libraries were sequenced to 40x coverage. T cells represent an ideal system for studying clonal populations, because each naive T cell expresses a unique T cell receptor (TCR) produced through genomic rearrangements of two separate regions of the genome located on Chromosomes 7 (β chain) and 14 (α chain). Activation of naive T cells triggers a rapid clonal expansion with each daughter cell inheriting the rearranged TCR genes. This allows the identification of clonally related cells by sequencing the genomic regions containing the TCR genes. Vaccination triggers the expansion of a small number of unique T cell clones that can be identified with a fluorescently labelled dextramers linked to HLA Class I proteins in complex with a single viral epitope[8]. We isolated cells responding to previously identified minor viral epitope (HLA-B7:RPIDDRFGL) which we had found exhibited a reduced diversity of responding cells relative the dominant HLA-A2 epitope[9,10]. Single antigen-specific $CD8^+$ T cells were isolated from longitudinal samples of peripheral blood by FACS and DNA was amplified by multiple displacement amplification (MDA) (**Fig. 1d**, **Supplementary Fig. 3**). High quality single cell MDA libraries were subsequently screened by PCR against the TCR α and/or β chains to identify clonal populations of T cells (**Fig. 1e, Online Methods**). Libraries from two clones (clones A and B), as well as two unrelated cells (UR) were subjected to whole genome sequencing (WGS) on the Illumina HiSeqX platform (**Supplementary Fig. 4)**. The 18 single cell libraries were analyzed by Conbase, and 117 putative sSNVs were identified when requiring that a variant was observed in at least 2

samples covered by at least 10 reads. Hierarchical clustering based on the sSNVs demonstrated unambiguous identification of each T cell clonal population (**Fig. 1f, Online Methods**).

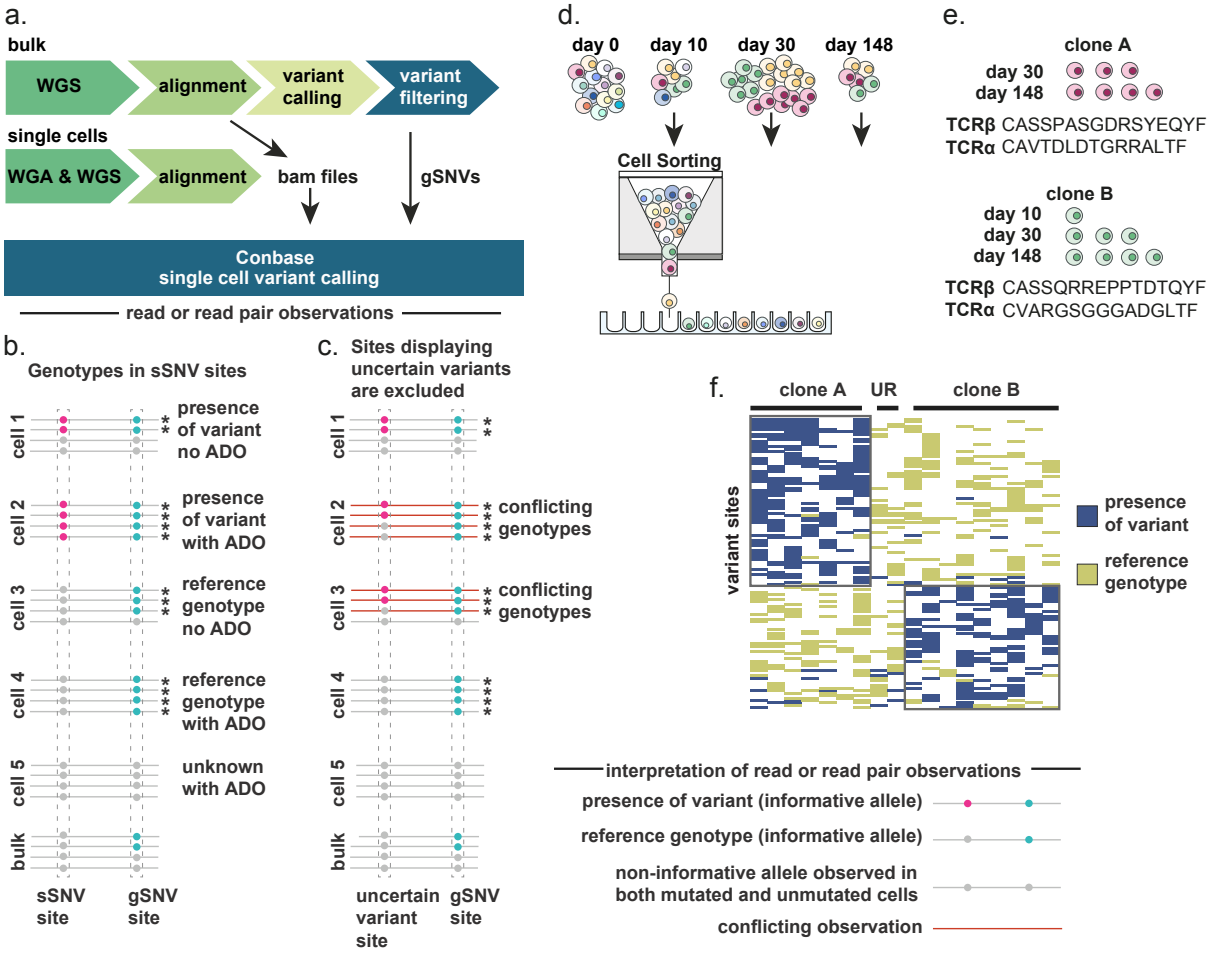


**Figure 1**. Identification of clonal sSNVs by Conbase. (a) Workflow depicting input data preparation. (b) True sSNVs are present on the same allele in mutated samples (cell 1, cell 2), representing the informative allele indicated by * and blue circle in the gSNV site. The informative allele is distinguished from the non-informative allele by the base observation in gSNV sites. Presence or absence of mutations can be determined despite ADO in samples displaying reads originating from the informative allele (cell 2, cell 4). The genotypes of samples only displaying reads from the non-informative allele are unknown (cell 5). (c) Sites displaying uncertain variants are excluded by monitoring the density of variant calls within a genomic window and by monitoring samples displaying support for more than two haplotypes, represented by conflicting observations with support for both mutated and unmutated genotypes in individual cells (cell 2, cell 3). Such observations may be the result of amplification errors affecting multiple cells, or observations present within low complexity genomic sequences, structural variants or genomic regions suffering from errors in the reference genome. (d-e)

Identification of single CD8+ T cells from clonally expanded populations taken longitudinally after vaccination of a healthy donor using TCR sequences and distributions of clonal T cells identified for sequencing. (f) Unsupervised hierarchical clustering of single CD8+ T cells from clones A and B and the two unrelated cells (UR) after variant calling (n=117 total variants used for clustering).

In the provided output from Conbase, predicted genotypes are color coded based on the presence or absence of the mutation, ADO status and read depth support (**Fig. 2a, Supplementary Fig. 2**). Validation of a selection of the identified mutations was performed by screening libraries that were identified by TCR sequencing as being related to the two T cell clones included in the study, as well as a third unrelated clone (Clone C), using PCR amplification of regions containing both sSNVs and gSNVs and Sanger sequencing of purified amplicons (**Fig. 2b**). Here we could confirm the presence of the identified mutations in additional cells identified as belonging to clones A and B, but never in clone C, demonstrating that predicted sSNVs represent true somatic mutations found throughout each clone. Importantly, this approach provides a useful platform for expanding the analysis to hundreds or thousands of cells using targeted screening after identification of high confidence mutations by Conbase.

We evaluated the fraction of predicted variant sites, which we could confidently determine to be false. These were defined as sites where a variant was observed in both clones, while at least one cell displayed the reference genotype. Such distribution of genotypes is biologically implausible, under the assumption that the probability of two mutations occurring independently in the same site twice is extremely low[12]. The fraction of sites displaying biologically implausible genotype distributions in the T cell clones was small, even at low read depth cut offs (**Fig. 2c**)

The ability of Conbase to call true somatic mutations, was compared to Monovar, a variant caller designed to specifically deal with the errors and biases present in single cell data[4]. Variant calling using Monovar resulted in a higher number of sites where multiple cells

shared a variant and at least one sample displayed the reference genotype, although the vast majority of these displayed implausible genotype distributions. (**Fig. 2c**). Applying more stringent filters resulted in fewer potential somatic variant sites, although these still displayed implausible genotype distributions (**Supplementary Fig. 5**). In positions that overlapped between Monovar and Conbase, we observed that Monovar incorrectly assigned reference genotypes to samples displaying dropout of the informative allele (**Supplementary Fig. 2**).

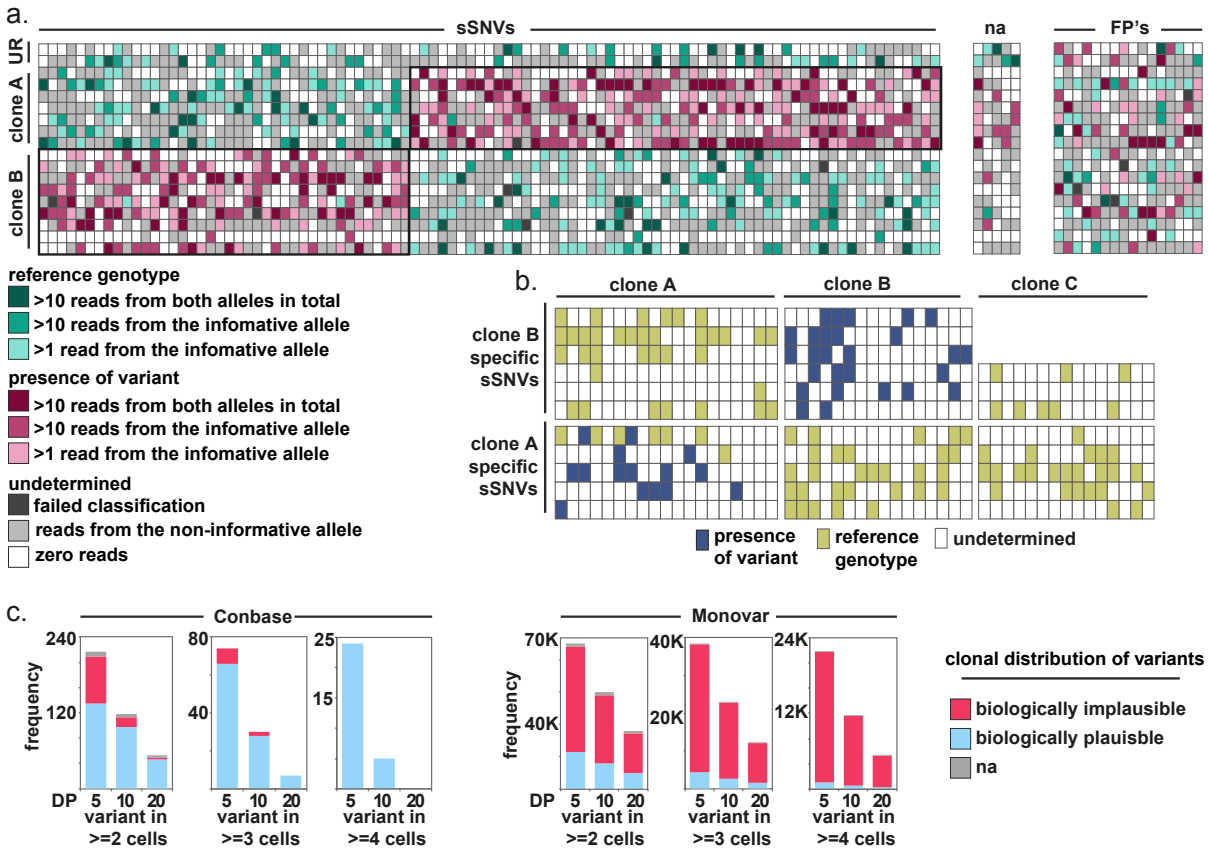


**Figure 2**. Conbase output, validation of selected variants in matched clonal cells, and comparison of Conbase with Monovar on T cell clones. (a) Expanded heatmap depicting variants used for clustering in figure 1c (n=117 total variants) showing statistics for reads used to identify variants at each position and distribution of false positive (FP, n=16) and undetermined (na, n=5) variant sites. (b) Sanger sequencing results for selected variants in amplified DNA from single CD8+ T cells identified by TCR sequencing as belonging to clones A, B, or an unrelated clone C. (c) Clonal distribution of variants predicted by Conbase or Monovar at different read depths (DP) and when observing the variant in increasing number of cells. Variants were identified as being biologically plausible or implausible using clonal distribution patterns and not available (na) if insufficient data existed to define clonal distribution.

Our philosophy in designing Conbase was to find a strategy to eliminate false variant calls in WGS data from amplified single cell DNA, in order to identify variants that can be used to screen additional cells for the presence or absence of the somatic mutations. Our results highlight the utility of Conbase as an algorithm for identifying high confidence mutations in single cell genome sequencing data from primary human cell populations isolated directly ex vivo. The fact that variant calling can be performed despite high rates of ADO opens up the possibility of analyzing cells while maintaining their spatial orientation in tissue through isolation by LCM, where the cryosectioning procedure may result in additional loss of genomic DNA prior to WGA. We believe that these features of Conbase will make it increasingly useful as single cell WGS becomes a standard practice in laboratory research and in medical diagnostics.

**Acknowledgements**

This work was supported by the Swedish Research Council, the Swedish Cancer Society, the Swedish Society for Strategic Research, Tobias Stiftelsen, AFA Försäkringar, the Strategic Research Programme in Stem Cells and Regenerative Medicine at Karolinska Institutet (StratRegen), Torsten Söderbergs Stiftelse and the Swedish Genomes Program (Science for Life Laboratory). The Swedish Genomes Program has been made available by support from the Knut and Alice Wallenberg Foundation. J.H. would like to acknowledge bioinformatics advice through the Swedish Bioinformatics Advisory Programme. Å.B. and B.S. are financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab.

**Author Contributions**

J.H. and J.E.M devised the project. J.H., M.K. and E.A.H. developed the software, M.K and E.A.H implemented and J.H analyzed data. J.E.M. conceived and planned the experiments with support from I.D., M.P. and J.M. Å.B., E.B., P.S., B.S. contributed to data analysis. J.H., M.K., E.A.H., and J.EM. wrote the manuscript with contributions from the remaining authors. J.E.M. and J.F. supervised the project.

**Competing financial interests**

The authors declare no competing financial interests.

References

1. Lodato *et al. Science.* **350**, 94-98 (2015)

2. Yu et al. *Nature.* **543**, 714–718 (2017)

3. Gawad, C., Koh, W., Quake, S.R. *Nature Reviews Genetics.* **17**, 175-188 (2016)

4. Zafar, M., Wang, Y., Nakhleh, L., Navin, N., Chen, K. *Nature Methods.* **13**, 505–507 (2016)

5. Dong *et al. Nature Methods.* **14**, 491–493 (2017)

6. Chen *et al*. *Science.* **356**, 189-194 (2017)

7. Li, H. *Bioinformatics.* **30**, 2843–2851 (2014)

8. Altman, J.D. *et al.* Science. **274**, 94-96 (1996).

9. Blom, K. *et al*. *J Immunol* **190**, 2150-2158 (2013)

10. Reinius, B *et al*. *Nat Genet* **48**, 1430-1435 (2016)

11. Zong, C., Lu, S., Chapman, A.R., Xie, X.S. *Science* **338**, 1622-1626 (2012)

12. Kimura M. *Genetics*, **4**, 893-903 (1969)

## Methods

### The Conbase algorithm

Conbase requires WGS data from WGA amplified single cells and an unamplified bulk sample to predict sSNV sites and genotypes in these positions. Conbase takes three inputs: single cell and bulk bam files, a human reference genome in fasta format and gSNV coordinates to be used for phasing. gSNV coordinates and gSNV base observations are obtained from vcf output previously generated from variant calling in a bulk sample by Freebayes (or another variant caller). The analysis is split up in two subprograms: stats and analyze. Stats outputs a json-file with unfiltered variant calls, which is used as input to analyze. The output of analyze are phased filtered variant calls summarized in a tsv-file, a pdf-file displaying color coded genotypes and an interactive html-file with detailed information about the predicted genotypes. Conbase variant calling is based on a number of assumptions associated with expected observations in true variant sites. These assumptions have been coded as adjustable parameters, with values that reflect how much the observations may deviate from the expected characteristics of a true variant site. The parameter values generating the data presented in the current report are specified in parentheses in the algorithm description below.

#### *stats*

The initial step consists of identifying sites supporting an alternative (potentially mutated) base as compared to the bulk sample. This is done by only considering bases present in the same read or read pair as at least one gSNV using a BAM reader (pysam). The genomic windows in which the analysis will take place are determined by defining the longest distance upstream and downstream of each gSNV, covered by read pairs within the dataset. Because of sequencing errors, amplification errors or misplaced reads, more than one alternative base

may be observed in the same position, despite a diploid genome. In order to determine the most probable alternative base in a site, we utilize the accumulative information given by all qualified samples in the following way:

Let $N_b^s$ be the read depth of base $b$ in sample $s$ at a given position.

$$R = reference\ base\ in\ bulk$$

$$b \in B = \{A, C, G, T\} \setminus \{R\}$$

Let $V_b^s$ define the voting result for a base $b$ by a sample $s$

$$V_b^s = \begin{cases} 1 \quad if\ \frac{N_b^s}{N_R^s} \geq \alpha, N_b^s + N_R^s \geq \beta \quad (1) \\ 1 \quad if\ \frac{N_b^s}{N_R^s} < \alpha, N_b^s \geq \beta \quad (2) \\ 0 \quad otherwise \end{cases}$$

where $\alpha$ (0.2) and $\beta$ (10x) are parameters for variant allele fraction and total read depth, determining which samples are qualified to vote. In the first case (1), reads from two alleles are observed, displaying sufficient variant allele fraction and total read depth. In the second case (2), the variant allele fraction is not sufficient, however the sample displays sufficient read coverage supporting the alternative base, which represents another indication for a potentially true variant since amplification bias may result in dropout of the reference base.

Let $V_b$ be the number of samples that voted for an alternative base $b$. If $V_b > 0$ we may define the most probable alternative base $A$ as

$$A = \begin{cases} argmax_b(V_b | \forall b \in B) & if\ \frac{V_b}{|V|} \geq \gamma,\ |V| \geq \varepsilon \\ undefined & otherwise \end{cases}$$

where $\gamma$ (0.9) and $\varepsilon$ (2) define the fraction of samples required to vote for the same base and the number of samples required to vote at a given position. $|V|$ is the number of samples that

were qualified to vote.

***analyze***

The previous processing identified potential sSNV sites in near distance of gSNVs where multiple cells shared strong support for the same alternative base $A$. If we consider the alternative base $A$ in the sSNV site in relation to the base observed in a gSNV site in the same read or read pair, we denote it as the following relation:

$$X_{mut}Y_{psnv}$$

where the sSNV site $X$ either displays the same base observed in the reference genome (and the bulk sample) or the alternative base $A$, and $Y$ either displays the same base as the reference genome or the alternative base in the gSNV site (obtained from vcf output from variant calling in the bulk sample). For simplicity, we will further denote these relationships as $RR, RA, AR, AA$ where the first letter refers to the observation in the sSNV site and the second letter refers to the observation in the gSNV site. From here on we will denote a given pair as a *tuple*.

Considering both the maternal and paternal pair, there are only two plausible couplings that could equal a true heterozygous genotype. Likewise, there is only one single type of coupling that constitutes a true homozygous genotype.

$$Heterozygous\ genotype = \{RR, AA\}, \{RA, AR\}$$

$$Homozygous\ genotype = \{RR, RA\}$$

Any of these three combinations is a combination of two tuples, and will further be referred to as a *tuplepair* ($tp$). Given that we expect a clonal mutation to be present in a subpopulation of the analyzed cells, we can collectively determine the expected coupling representing a true heterozygous genotype, either $\{RR, AA\}$ or $\{RA, AA\}$, by analyzing tuplepairs in all the samples that qualifies for voting.

All samples that display reads covering the tuple (i.e. the reads contain the specific gSNV in mind and the sSNV) with a total read depth $> \nu$ (10x) are allowed to vote if they fulfill the following criteria: Let $tp_{max}, tp_{min}$ be the tuplepairs with the highest and lowest depth out of the two possible tuplepairs for heterozygous positions. A sample is allowed to vote for $tp_{max}$ if i) the ratio between the read depth of $tp_{max}$ and $tp_{min} < \rho$ (0.01) and ii) either the ratio for the tuples in $tp_{max} > \lambda$ (0.1) or if the tuple covering the variant $\{AR\ or\ AA\}$ is the maximum out of all tuples for that sample. Out of the entitled samples that voted for their local $tp_s$, the most probable tuplepair across all samples $tp^*$ is finalized if at least $\kappa$ (2) samples voted and if $tp^*$ held a majority of at least $\omega$ % (90%) of the votes. If the voting samples do not manage to conclude the final tuplepair for a sSNV in connection to a specific gSNV, that will dictate across all samples, the allelic origin of the alternative base is simply unknown.

By knowing the designated $tp^*$ we also know which of the two alleles we expect to observe the variant on, either $\{AA\}$ or $\{AR\}$. Consider the following: while we may jump to the conclusion of having found a variant as soon as we observe e.g. tuple $AA$ in a sample, if the $tp^*$ actually consists of tuples $\{AR, RA\}$ we actually know for certain that this cannot be a true variant as such observation would be contradictory. Moreover, given this $tp^*$, if a sample only display reads supporting the tuple $RA$ we obviously cannot determine the genotype since RA is expected to be observed in both heterozygous and homozygous genotypes.

Let $N_x$ be the read depth of the tuple $x$ and $N_{tot}$ the total read depth for all tuples. Consider the case when $tp^* = \{RR, AA\}$ and let

$$d_I^{het} = \frac{min(N_{RR}, N_{AA})}{N_{RR} + N_{AA}}$$

$$d_E^{het} = \frac{N_{RR} + N_{AA}}{N_{tot}}$$

$$d_I^{hom} = \frac{min(N_{RR}, N_{RA})}{N_{RR} + N_{RA}}$$

$$d_E^{hom} = \frac{N_{RR} + N_{RA}}{N_{tot}}$$

where $d_I^x$ is the (internal) ratio for the tuples and $d_E^x$ is the (external) ratio for a given tuplepair in relation to all of the tuples for a site from a homozygous and heterozygous perspective. These ratios will be used to determine the confidence in the allelic origin of the sSNV for a given site. The case when $tp^* = \{RA, AR\}$ is treated analogously.

If $tp^*$ is $\{RR, AA\}$, reads displaying $A$ in the gSNV site is the informative allele, and reads originating from the informative allele are required for genotyping. If $tp^*$ is $\{RA, AR\}$, reads displaying $R$ in the gSNV site is the informative allele. Samples that display reads originating from both alleles are confidence level 1 genotypes (C1) and samples that display reads originating from the only the informative allele are confidence level 2 genotypes (C2). A genotype determination $GT_{SNV}^s$ for C1 and C2 for a given position in relation to a specific gSNV for a sample $s$ will only be possible if there is a given $tp^*$ and total read depth of all tuples > $\varphi$ (10x). If these initial conditions are met, $GT_{gSNV}^s$ is computed as follows:

$$GT_{gSNV}^s = \begin{cases} het_{C1} & if\ d_E^{het} \geq \theta, d_I^{het} \geq \tau \\ hom_{C1} & if\ d_E^{hom} \geq \theta, d_I^{hom} \geq \tau \\ het_{C2} & if\ d_I^{het}, d_I^{hom} > \tau, d_E^{het} \geq \theta, d_E^{hom} < 1 - \theta \\ hom_{C2} & if\ d_I^{het}, d_I^{hom} < \tau, d_E^{hom} \geq \theta, d_E^{het} < 1 - \theta \\ conflict & if\ conditions\ met\ for\ het_{C1}, \mathrm{hom}_{\mathrm{C1}} \\ conflict & if\ tp_s \neq tp^* \end{cases}$$

where $\tau$ (0.1) is the internal ratio parameter and $\theta$ (0.9) is the external ratio parameter.

The reads covering the sSNV may cover multiple gSNVs, which is why a collectively decided final genotype can be defined by letting all qualified $GT^s_{SNV}$ vote for the final genotype. $GT^s_{max}$ is therefore the genotype with the most votes. Let $N_{max}$ be the number of gSNVs that voted for $GT^s_{max}$, $N_{vot}$ is the number of gSNVs that were qualified to vote ($tp^*$ is defined for this gSNV). Let $N_{tot}$ be the number of gSNVs that were present in the same read or read pairs as the sSNV and let $GT^s$ be defined as

$$GT^s = \begin{cases} GT^s_{max} & if \quad \frac{N_{max}}{N_{vot}} \geq \psi, \frac{N_{vot}}{N_{tot}} \geq \sigma \\ conflict & if \quad \frac{N_{max}}{N_{vot}} < \psi \end{cases}$$

where $\psi$ (0.9) is the fraction of gSNVs voting for the same genotype, $\sigma$ (0.9) is the fraction of gSNVs qualified to vote for a genotype.

If $N_{vot} = 0$, any available read information is used to infer genotypes for samples displaying low read depth, albeit to a lower degree of confidence (confidence level 3 (C3)). Support for an unmutated genotype is defined as when either the $tp^* = \{AA, RR\}$ and the sample displays a read depth supporting the tuple $RA > \chi$ (1x) or when the $tp^* = \{AR, RA\}$ and the sample displays a read depth supporting the tuple $RR > \chi$, resulting in $GT^s = hom_{C3}$. If there is no support for an unmutated genotype and the sample displays a read depth supporting the alternative base $> \xi$(1x), $GT^s = het_{C3}$. Samples displaying support for both a mutated and an unmutated genotype are considered $GT^s = conflict$.

***Conflict-filtering***

Samples classified as $GT^s = conflict$ are often present within sites or in proximity to sites where other samples display conflicting genotypes and biologically implausible genotype distributions (**Supplementary Fig. 2**). Excluding sites where any sample display conflicting

genotypes filters out many false variant sites. In the current report zero samples displaying conflicting genotypes were allowed per site.

***Cluster-filtering***

A common artefact in variant calling output, is regions with clusters of false positive mutations, correlating with areas in the genome regions with poor mappability and repetitive elements. In the current report, a maximum of 1 mutation per kb was allowed.

**Sequence alignment and data processing**

Following WGS and demultiplexing, the reads were trimmed from Illumina adapters using Cutadapt[13]. WGA adapters were trimmed from MALBAC amplified samples using Cutadapt[13]. Read pairs were aligned to the human genome (human g1k v37 with decoy) using Burrows-Wheeler aligner (BWA-MEM)[14]. Processing of the mapped reads and sequencing data quality evaluation was performed using Picard Tools (http://picard.sourceforge.net) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read processing included removal PCR duplicates, optical duplicates and reads with a mapping quality below 2, including multimappers. Indel realignment was performed with GATKs IndelRealigner[15]. Variant calling was performed using Freebayes with default settings[16].

***gSNV filtering***

Following variant calling in bulk samples using Freebayes, variants were filtered by vcffilter (https://github.com/vcflib/vcflib) to identify gSNVs (**Supplementary Fig. 6**). False gSNVs will result in false variant calls in the downstream analysis with Conbase. As such, we applied stringent filters on the vcf output from Freebayes. With an average read depth of ≈40x in our bulk sample, we estimated a conservative maximum depth threshold of 55x at any position, as

$$d + 2.5\sqrt{d},$$

where $d$ is the average read depth across the genome.

We filtered variants on autosomes with a 15-55-fold read depth, quality score above 10 (QUAL), reads originated from both strands (SAF & SAR), at least two reads were balanced on each side of the site (RPR & RPL) and the alternative allele observation count was required to range between 20%-80% (AO). The variants were further filtered to remove gSNVs present in suspected erroneous regions in the reference genome. This was done by running a separate script included in the Conbase package, which screens the bulk bam file around each gSNV, and excludes gSNVs present within a 1kb window containing >10 'heterozygous' positions. A 'heterozygous' position was defined as a position where >10% of the reads supported a non-reference base. 1789830 gSNVs were identified in unamplified bulk peripheral blood cell sample used as reference for Conbase.

**Clonal T cell Isolation and Analysis**

***Sample Preparation and Cell Sorting***

Study participants were recruited into an ongoing study to monitor immune responses to the yellow fever virus vaccine YFV-17D (approved by the Regional Ethical Review Board in Stockholm, Sweden: 2008/1881-31/4, 2013/216-32, and 2104/1890-32). A female subject was identified based on being positive for HLA-B7 and having a detectable T cell response to a minor peptide (RPIDDRFGL) presented by HLA-B7. Cryopreserved peripheral blood mononuclear cell (PBMC) samples taken at days 10, 30, and 148 post-vaccination were thawed at 37°C and quickly washed in FACS buffer (PBS with 2% BSA/2mM EDTA). Negative selection with magnetic beads was performed for each sample to purify $CD8^+$ T cells (Miltenyi Human CD8 Negative Selection kit, 130-096-495). Purified $CD8^+$ T cells were first incubated with an HLA-B7/RPIDDRFGL dextramer conjugated to Alexa fluor 647 (Immudex) for 15 minutes. Cells were subsequently incubated with a panel of antibodies to

identify live $CD3^+CD8^+$ T cells (CD3–Alexa Fluor 700 (UCHT1, BD Biosciences), CD8-APC-Cy7 (SK1, BD Biosciences), CD4-PE-Cy5 (RPA-T4, eBioscience), CD14–Horizon V500 (MΦP9, BD Biosciences), CD19–Horizon V500 (HIB19, BD Biosciences), and Live/Dead Fixable Aqua Dead Cell staining kit (Invitrogen, L34957)). Live, Lineage negative $CD3^+CD8^+Dextramer^+$ cells were sorted into 96 well PCR plates (Thermo Scientific, AB-0800) containing lysis buffer (200mM KOH, 40mM DTT, 5mM EDTA). Single cells were incubated on ice for 10 minutes in lysis buffer after which neutralization buffer (400mM HCL, 600mM Tris-HCL pH7.5) was added followed by an additional 10-minute incubation on ice. Lysed cells were subsequently stored at -80°C until amplification reactions were performed.

***WGA by MDA***

Lysed single cells were subjected to multiple displacement amplification (MDA) as previously described[17]. A mixture containing dNTPs (Invitrogen, 2mM), random hexamer primers with 3' thiophosphate linkers (5'-dNdNdNdN*dN*dN-3', IDT (50uM)), and repliPHI polymerase (40U) in phi29 reaction buffer (Epicentre) was added to each well to bring total volume to 20uL. Cells were incubated at 30°C for 10 hours followed by a 3-minute incubation at 65°C to inactivate the phi29 polymerase. The resulting libraries were diluted in $H_2O$ to 50uL and concentrations of double stranded DNA were measured (Qubit, Broad Range dsDNA kit).

***Identification T Cell Receptors from Single Cell MDA Material***

We adopted a previously published method[18] so that we could screen large numbers of single cell libraries to identify clonally related T cells by TCR rearrangements. Approximately 100ng of amplified DNA was taken from each sample and touchdown PCR (Tm: 72°C ->

55°C) was performed using a panel of primers designed upstream of each variable region and downstream of the joining regions for the human TCR α or β chain locus. A dilution of each reaction was subsequently used to perform a second, nested-touchdown PCR with internal primers designed against each variable and joining region of the human TCR α or β chain locus. The internal primers contained handles which were used to index each well for the 96-well plate so that they could be pooled into a single reaction. Each plate was then prepared according to the Truseq (Illumina) protocol for sequencing on an Illumina Miseq (2x150bp reads).

After demultiplexing of Illumina sample indexes, the reverse read (R2, 150bases) Fastq file was converted to Fasta format. Identical sequences were clustered using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) FASTA Collapser. Then sequences were sorted by our 96 well indexes using the FASTX barcode splitter, and the first 44 bases were finally trimmed off using the FASTA trimmer to facilitate downstream sequence analysis. Because the internal primers targeting the joining regions were within 50bp of the CDR3 region of the TCR it was possible to identify clonal T cells based on shared CDR3 nucleotide sequences. All samples were individually analyzed using the IMDT database to identify the CDR3 sequence.

***Selecting High Coverage Libraries for Illumina Sequencing***

Clonal T cells were grouped and high quality libraries were identified using a panel of chromosome specific PCR primers as described previously[11]. High quality T cell libraries were considered to be samples with detection at the majority of loci and were subsequently processed using a PCR-free TruSeq library preparation kit (Illumina) and sequenced with a HiSeq X using a theoretical coverage of 30x per sample (SciLifeLab, Karolinska Institute).

***Screening Related Clonal T Cells by Sanger Sequencing***

Single cell libraries that were included in the original screening which matched clones A or clone B were identified to be used for verifying selected mutants. An additional clone (Clone C (TCRα: CAAHSPYSGNTPLVF, TCRβ: CASSSGTAYNEQFF) was used as a control to determine whether mutations could be found as artifacts in unrelated T cells. Primers were designed to span both the gSNV and the putative variants and samples were subjected to 35 cycles of PCR (Tm: 67°C) (PCRBIO HiFi Polymerase, PCR Biosystems) yielding approximately 1000bp amplicons. Additionally, primers contained handles (similar to those used for TCR screening) so that secondary amplification cycles could be used to index samples if necessary. Amplified samples were analyzed by gel electrophoresis and bands were excised for DNA isolation (Nucleospin Gel Clean Up, Techtum). Gel-purified DNA samples were sent for Sanger sequencing (KI Gene Facility, CMM, Karolinska Institute) using primers specific for the universal handle incorporated onto each Forward primer. Sanger sequencing results were analyzed visually using the software package 4peaks.

**Comparisons of single cell variant calling algorithms for performance evaluation**

Monovar[4] was run on single T-cell amplified with MDA with default settings (ADO was set to 90%) and consensus filtering. From the raw Monovar output, sites present within 10bp of another site were removed. Sites overlapping with raw variants called in an unamplified bulk sample by Freebayes, were removed. Potential somatic variants were filtered on autosomes by requiring that at least two samples shared a variant (0/1 or 1/1) while at least one sample displayed the reference genotype (0/0). Samples in the same site which did not pass these cut offs were assigned an unknown genotype. The fraction of sites displaying biologically plausible and biologically implausible genotype distributions was computed (**Fig. 1c**, **Supplementary Fig. 5**). Biologically plausible genotype distributions were defined as sites

were at least one cell displayed a reference genotype and multiple cells within clone A or clone B shared variant. Biologically implausible genotype distributions were defined as sites where at least one cell display the reference genotype, and at least one cell displayed a mutated genotype (0/1 or 1/1) in both clones. Whether the distribution of genotypes were biologically plauisble or not was undetermined in sites where the genotypes of all cells within one clone were unknown, due to zero reads or genotypes not passing filters. Variants were parsed from Monovar output, at different genotype quality cut offs (GQ), read depth cut offs (DP) and by requiring that the variant was observed in increasing number of cells (**Fig. 1**, **Supplementary Fig. 5**).

We attempted to call variants using SCcaller[5]. The rate of amplification bias observed in the T cell dataset resulted in eta-values that were too low to enable distinction between true mutations and artefacts (personal communication with Xiao Dong[5]). Thus, we did not move forward with variant calling using the SCcaller.

**Hierarchical clustering**

Variants called by Conbase were used to define phylogenetic distances between cells.
Distances between cells were defined as unknown if no shared sites were detected.
For shared sites, the distance was decreased with -1 for each site where cells have the same call (mutated or non-mutated) and increased with +1 for sites where the cells have different calls. The distance matrix was then clustered using standard hclust with the distance 'ward.D2' (**Fig. 1f**).

**Software**

Conbase is implemented in python and is only dependent on the pysam module.

**References**

13. Martin, M. *EMBnet* **17**, 10-12 (2011)

14. Li H, Durbin R. *Bioinformatics* **14**, 1754-60 (2009)

15. McKenna, A. *et al. Genome Res* **20**, 1297-1303 (2010)

16. Garrison, E. & Marth, G. arXiv:1207.3907 (2012)

17. Evrony, G.D. *et al*. *Cell* **151**, 483-496 (2012).

18. Balazs, A.B., Tsai, J.M., Baltimore, D. **Patent (US8497071 B2)** (2013)