Ultra-Fast Local-Haplotype Variant Calling Using Paired-end DNA-Sequencing Data Reveals Somatic Mosaicism in Tumor and Normal Blood Samples

Subhajit Sengupta[1], Kamalakar Gulukota[2], Yitan Zhu[1],
Carole Ober[4], Katherine Naughton[4], William Wentworth-Sheilds[4], Yuan Ji[1,3] *

Abstract

Somatic mosaicism refers to the existence of somatic mutations in a fraction of somatic cells in a single biological sample. Its importance has mainly been discussed in theory although experimental work has started to emerge linking somatic mosaicism to disease diagnosis. Through novel statistical modeling of paired-end DNA-sequencing data using blood-derived DNA from healthy donors as well as DNA from tumor samples, we present an ultra-fast computational pipeline, *LocHap* that searches for multiple single nucleotide variants (SNVs) that are scaffolded by the same reads. We refer to scaffolded SNVs as local haplotypes. When a local haplotype exhibits more than two genotypes, we call it a local haplotype variant (LHV). The presence of LHVs is considered evidence of somatic mosaicism because a genetically homogeneous cell population will not harbor LHVs. Applying LocHap to whole-genome and whole-exome sequence data in DNA from normal blood and tumor samples, we find wide-spread LHVs across the genome. Importantly, we find more LHVs in tumor samples than in normal samples, and more in older adults than in younger ones. We confirm the existence of LHVs and somatic mosaicism by validation studies in normal blood samples. LocHap is publicly available at `http://www.compgenome.org/lochap`.

Keywords: Evolution; LocHap; NGS; SNV; Somatic mosaicism; Somatic mutations;

INTRODUCTION

Many cancers arise from a series of mutational events occurring throughout a person's life span [1, 2]. Considerable evidence [3, 4] has accumulated supporting the presence of genetically heterogeneous cells in a somatic sample, a phenomenon called somatic mosaicism, which may be a precursor to the onset of many cancers [5]. However, there are no effective and economical tools that can reliably measure the presence and degree of somatic mosaicism in a biological sample. Single cell sequencing [6] in principle provides the

*[1]Program of Computational Genomics & Medicine, NorthShore University HealthSystem, Evanston, IL, USA. [2]Center for Molecular Medicine, NorthShore University HealthSystem, Evanston, IL, USA. [3]Department of Health Studies, University of Chicago, Chicago, IL, USA. [4]Department of Human Genetics, University of Chicago, Chicago, IL, USA. Correspondence should be addressed to Y.J. (koaeraser@gmail.com).

genetic landscape of each individual cells, although in practice only up to hundreds or thousands of cells can be measured due to the formidable cost of money and effort. In contrast, next-generation sequencing (NGS) technologies assemble an average genome sequence of all the cells in a sample, assuming cellular homogeneity. In the presence of somatic mosaicism, the average genome may not be a good representation of the sample. Despite continuous breakthroughs in DNA sequencing since the completion of the human genome project [7], researchers are still unable to precisely dissect individual cellular genomes on large scales.

Somatic mosaicism is often seen in samples derived from patients with cancer. Future targeted and personalized cancer therapy must take into account mosaic tumor cells in order to better customize therapies [8, 9]. In contrast, somatic mosaicism in samples from healthy individuals has been discussed as a theory over the last decade [10–12], with only a few recently reported examples [5, 13–20]. Due to the availability of high-throughput DNA sequencing, hundreds of millions of short reads can now be mapped to cover whole genomes or exomes. If somatic mosaicism is present in a biological sample, the DNA sequences of the short reads are expected to reflect the variations of the cellular genomes at the single nucleotide level. Based on this concept, pioneering work in 2014 by Genovese et al. [5] reported the presence of somatic mutations in blood samples as precursor of hematologic cancer and death. They carefully constructed bioinformatics and statistical methods to filter single nucleotide variants (SNVs) based on whole-exome and whole-genome sequencing data and identified clonal somatic blood samples with somatic mutations. Because the somatic mutations were only present in a fraction of the cells, the blood sample was considered mosaic. Their main computational analysis aimed to identify SNVs with variant allele fractions (VAFs) that are far smaller than 0.5 and attributed these SNVs to the existence of small cellular subpopulations harboring the SNVs. Computationally it is challenging to differentiate true biological subpopulations from noise and artifact in the NGS data since both would give rise to small VAFs [21].

We propose here a different approach. Instead of using SNVs, we consider **local haplotypes** (LHs) for calling somatic mosaicism. A LH is a scaffold of multiple proximal SNVs (Fig. 1). Examining paired-end DNA-sequencing data, we find that sometimes multiple SNVs are simultaneously mapped by the same short reads. The short reads provide linked genotypes for the SNVs. In Fig. 1, two SNVs are considered in each example and some short reads cover both SNVs. Treating the scaffold of the two SNVs as an LH, shown in Fig. 1, we observe three different genotypes with substantial read counts in each example. We call such a LH a local haplotype variant (LHV). The presence of LHVs across the genome is direct evidence supporting mosaicism and cellular heterogeneity because a homogeneous cell population can only manifest up to two haplotypes. Therefore, the key idea of examining an LH instead of an SNV allows for direct observation of more than two alleles in local genomes, a rare event for single loci but not for haplotypes. Based on this idea, we develop an open-source, ultrafast, and powerful computational tool, **LocHap**, for identifying LHVs using deep DNA-sequencing data from a single biological sample. We construct rigorous statistics models that provide probability measure for the LHVs. We also introduce bioinformatic filters that account for the usual noise and artifact in NGS data. However, the noise and artifact are partially mitigated due to the use of LHVs instead of SNVs. We elaborate more on these points in the next section. LocHap can be applied to any DNA-sequence data using paired-end reads and only requires a binary alignment and mapping (*bam*) file, the associated index (*bai*) file, and the corresponding variant call format (*vcf*) file (Materials and Methods). These files are almost always generated from standard variant-calling pipelines. To facilitate downstream analyses and experimental validation, we introduce a new file format, the haplotype call format, or *hcf*, that contains a list of LHVs inferred by LocHap. An *hcf* file has a tab-delimited format similar to a *vcf* file, and can be viewed in popular visualization tools like IGV [22]. The proposed *hcf* format is derived from the *vcf* format to facilitate visualization and interpretation. However, unlike *vcf* which contains SNVs and other genetic variants, *hcf* only contains information about LHVs, which is a scaffold of multiple local SNVs (each SNV is in the *vcf* file for the sample). Therefore, a non-empty *hcf* file presents information supporting genetically heterogeneous samples.

MATERIALS AND METHODS

Main Idea

The basic idea of LHV calling is to probabilistically model short reads mapped to multiple proximal SNVs and look for multi-allelic loci. In other words, we search for proximal SNVs that are scaffolded by short reads and exhibit more than two alleles with high statistical confidence. For example, Fig. 1a shows an LHV consisting of two SNVs, at chromosomal locations separated by only 97 base pairs. Examining data "horizontally" across both SNVs, many reads scaffold the SNVs as they are mapped to both loci. There are three directly observed haplotypes, GG, GC, and AG, with read frequencies 23, 10, and 9, respectively. In addition, four other types of overlapping short reads cover only one of the two SNVs. Each type of short reads potentially supports the presence of one or two different haplotypes and collectively they provide information on how many and what haplotypes are present in the region. Using all the short reads, LocHap employs a Bayesian hierarchical model, performs statistical inference accounting for the noise in the data, and filters dubious LHV calls based on false discovery rates (FDR) [23, 24].

Statistical Methods

SNV Segments

LocHap uses DNA-Seq data and assumes that base calling, reads alignment, and variant calling have been completed and *bam*, *bai*, and *vcf* files are available for one or more samples. LocHap first constructs non-overlapping segments on the genome, each of which is a set of continuous base pairs (bps) and contains at least two proximal SNVs separated by no more than K bps apart. The segment is formed by starting at a SNV, and extended to the next closest SNV as long as it is within K bps from the previous SNV. The segment ends if the next closest SNV is more than K bps away. Therefore, each segment starts and ends at a SNV, with potential multiple SNVs in between. A schematic illustration of DNA segmentation is shown in Fig. 2.

Along the genome, we start with the first called SNV, and form as many segments as we can until we reach the last called SNV. LocHap allows any integer K set by users as the maximum distance between two adjacent SNVs. For short-read data, we allow K to vary between 50 and 1,000. Changing K values will affect the size and number of segments. Usually the value of K can be set to reflect the insert length of the DNA sequencing experiment.

Probability Model for LHV Calling

LocHap analyzes each DNA segment separately. The goal of the analysis is to estimate the number and sequences of the haplotypes within the segment. Assume N numbers of short reads are mapped to the segment and each read overlaps with at least one SNV in the segment. Mapped reads that do not overlap with any SNVs are discarded since they do not contribute to the haplotype calling.

For a given segment, let $i = 1, \ldots, N$ be the index of the mapped short reads. Assume that R SNVs are present in the segment. We consider up to $L = 2^R$ candidate haplotypes that can be formed by R SNVs. That is, we assume at each SNV, one can observe up to two different alleles (e.g., a reference and a variant allele). More than two alleles are rarely observed from short reads for an SNV and most of them are caused by sequencing error. We use $j = 1, \ldots, K$ to index possible haplotypes. The genotypes (nucleotide sequences) of each candidate haplotype are denoted by $\boldsymbol{h}_j = \{h_{j1}, \ldots, h_{jR}\}$, $j = 1, \ldots, L$, where h_{jr} takes one of the four nucleotides, i.e., $h_{jr} \in \{A, C, G, T\}$ for $r = 1, \ldots, R$. For example, in Fig. 1, $R = 2$ and $L = 2^R = 4$. In each example some short reads overlap with both SNVs. Specifically, for each short read i, use $\boldsymbol{s}_i = \{s_{i1}, \ldots, s_{iR}\}$ to denote an R-base DNA sequence of interest, where $s_{ir} \in \{A, C, G, T, M\}$; here M denotes a missing base readout when there is no overlap between a short read and an SNV. Let $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ be the set of all short reads. One could define an indicator $m_{ir} = I(s_{ir} = M)$ to denote the missing base of $\boldsymbol{s}_i$ and set up a model $f(m_{ir} \mid \theta)$. We assume missing completely at random (MCAR) [25] , which leads to conditional independence in the posterior inference. That is, conditional on $\boldsymbol{s}$, parameters in the model describing target haplotypes are independent of $\boldsymbol{m} = \{m_{ir},\ \ i = 1, \ldots, N, r = 1, \ldots, R\}$, the vector of missing indicators. This greatly simplifies the inference procedure. The MCAR assumption is proper here since in NGS experiment, typically the missing base in s_{ir} is due to that read i is not aligned to base SNV r, which is caused by the limited read length as a technological limitation. Hence the missing mechanism in s_{ir} has nothing to do with what sequences are observed or not observed.

Using standard missing data notations, let

$$\boldsymbol{s}^{obs} = \{s_{ir}, \quad \text{where } m_{ir} = 0 \text{ for } i = 1, \ldots, N;\ r = 1, \ldots, R\}$$

and

$$\boldsymbol{s}^{mis} = \{s_{ir}, \quad \text{where } m_{ir} = 1 \text{ for } i = 1, \ldots, N;\ r = 1, \ldots, R\}$$

denote the observed and missing DNA sequences for reads i at SNV r, respectively, for all i's and r's. Then $\{\boldsymbol{s}^{obs}, \boldsymbol{s}^{mis}, \boldsymbol{m}\}$ are the complete data, and $\{\boldsymbol{s}^{obs}, \boldsymbol{m}\}$ are the observed data. We introduce a few additional notations needed for modeling. Denote $\{\lambda_j = 1\}$ or $\{\lambda_j = 0\}$ the event that haplotype j is present or absent in the sample, respectively. Apparently λ_j's are key parameters of interest. Intuitively, the sequence similarity between haplotype sequences $\boldsymbol{h}_j$ and short read sequences $\boldsymbol{s}_i$ provides information on which haplotype is present. For example, if $\boldsymbol{s}_i$ matches $\boldsymbol{h}_j$ in most of the R bases, it is likely $\boldsymbol{s}_i$ is generated from a DNA segment having haplotype j, thereby supporting the presence of the haplotype. To model the similarity, we denote $\mathcal{A}_j(\boldsymbol{s}_i^{obs})$ and $\mathcal{D}_j(\boldsymbol{s}_i^{obs})$ the set of agreeing and disagreeing bases between $\boldsymbol{s}_i$ and $\boldsymbol{h}_j$, respectively. Mathematically, they refer to

$$\mathcal{A}_j(\boldsymbol{s}_i^{obs}) = \{r : s_{ir} = h_{jr}\}; \quad \mathcal{D}_j(\boldsymbol{s}_i^{obs}) = \{r : s_{ir} \neq h_{jr} \;\&\; s_{ir} \neq M\}.$$

Denote $I()$ the indicator function and let

$$\mathcal{B}_i = \{r : s_{ir} = M\} \text{ and } w_i = |\mathcal{B}_i| = \sum_{r=1}^{R} I(s_{ir} = M)$$

be the set of indices and number of missing bases of read i, respectively.

We propose a Bayesian probability model treating $(\boldsymbol{s}^{obs}, \boldsymbol{m})$ as observed data and $\{\boldsymbol{s}^{mis}, \lambda_j\}$ as unknown parameters. The inference is based on posterior probability that a haplotype j is present in the sample, $Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}, \boldsymbol{m})$. The higher value the probability takes, the more likely haplotype j is present. We will show next that this posterior probability can be calculated in a closed form.

Let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_L\}$ and $\boldsymbol{\lambda}_{-j} = \{\lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_K\}$ be the vector without the j-th component. The posterior probability $Pr(\lambda_j = 1 \mid \boldsymbol{s})$ can be calculated as follows.

$$\begin{aligned} Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}, \boldsymbol{m}) &= Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}, \cancel{\boldsymbol{m}}) \propto \underbrace{p(\boldsymbol{s}^{obs} \mid \lambda_j = 1)}_{\text{likelihood}} \underbrace{Pr(\lambda_j = 1)}_{\text{prior}} \\ &= \sum_{\boldsymbol{\lambda}_{-j} \in \mathcal{Y}_{L-1}} p(\boldsymbol{s}^{obs} \mid \lambda_j = 1, \boldsymbol{\lambda}_{-j}) \, Pr(\lambda_j = 1, \boldsymbol{\lambda}_{-j}) \\ &= \sum_{\boldsymbol{\lambda}_{-j} \in \mathcal{Y}_{L-1}} \left[\prod_{i=1} \underbrace{p(\boldsymbol{s}_i^{obs} \mid \lambda_j = 1, \boldsymbol{\lambda}_{-j})}_{\mathcal{I}} \underbrace{Pr(\lambda_j = 1, \boldsymbol{\lambda}_{-j})}_{\mathcal{II}} \right], \end{aligned} \tag{1}$$

where $\mathcal{Y}_{L-1}$ denotes the set of all binary (0 or 1) strings of length $(L-1)$. The first equation is due to the MCAR assumption. It can be shown (Supplementary Data) that

$$\begin{aligned} \mathcal{I} = \sum_{j'=1}^{L} \Bigg[& I(\lambda_{j'} = 1) \cdot c_{1i}(\lambda) \frac{1}{\sum_{\tilde{j}=1}^{L} \lambda_{\tilde{j}}} \times \\ & \sum_{\boldsymbol{b}_i \in \{A,C,G,T\}^{w_i}} \left\{ \prod_{r \in \mathcal{A}_{j'}(\boldsymbol{s}_i^{obs}, \boldsymbol{s}_i^{mis} = \boldsymbol{b}_i)} (1 - e_{ir}) \times \prod_{r \in \mathcal{D}_{j'}(\boldsymbol{s}_i^{obs}, \boldsymbol{s}_i^{mis} = \boldsymbol{b}_i)} \frac{e_{ir}}{3} \right\} \Bigg], \end{aligned}$$

where e_{ir} is the error probability for the DNA sequence called at base r on short read i. Typically e_{ir} is known from upstream analysis, e.g., in the form of *Phred quality score.* LocHap requires user-assigned values for e_{ir} with a default value of 0.001 (corresponding to a Phred score of 30). Alternatively, we recommend setting $e_{ir} = 10^{-logPh}$ where Ph is the Phred score at the base r of read i [26].

Next, the second term ($\mathcal{II}$) of Equation (1) is the product of independent prior term for each λ_j for all $j = 1, \cdots, L$,

$$\lambda_j \sim \text{Beta-Bernoulli}(\alpha, \beta, n), \quad \text{where } \lambda_j \in \{0, 1\}.$$

The beta-Bernoulli prior for λ is the marginal density of hierarchical construction in which

$$\lambda \mid \tau \sim \text{Bernoulli}(1, p); \qquad \tau \sim \text{Beta}(\alpha, \beta).$$

Integrating out τ, we get a beta-Bernoulli prior given by

$$Pr(\lambda = 1) = \frac{\Gamma(1+\alpha)\Gamma(\beta)}{\Gamma(1+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}. \tag{2}$$

To reflect the weak prior belief that a random haplotype has a low prior probability to be present in a sample, we set $\alpha = 0.05$ and $\beta = 1$ so that *a priori* the probability that haplotype j is present is only 5%.

FDR-based Inference and Calibration of e_{ir}

Denoting $\xi_j = Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}, \boldsymbol{m})$ the posterior probability that haplotype j is present in the sample. Posterior inference is based on selecting the haplotypes with the largest ξ_j subject to an FDR threshold. For example, with a desired FDR threshold of f_0, compute

$$j* = \max \left\{ j : \frac{\sum_{k<j}(1 - \xi_{(k)})}{|\{k : k < j\}|} < f_0 \right\} \tag{3}$$

where $\xi_{(k)}$ is the ordered statistics with decreasing order and $|\{set\}|$ is the cardinality of the set. Then select all the haplotypes with $\xi_j > \xi_{j*}$. Such a selection procedure is optimal [23,24] in controlling posterior expected FDR.

All the parameters in the proposed Bayesian model are estimated directly. The models only depend on one calibration parameter, e_{ir}, which must be given. The error rate e_{ir} captures the quality and Phred quality score from base calling, an upstream analysis. In most cases, a Phred quality score of > 30 is considered of high quality for a base, which translates to $e_{ir} < 0.001$ by definition (`http://en.wikipedia.org/wiki/Phred_quality_score`). Also, shown in Ji et al. [26] a higher error rate leads to more noisy inference, in our case, less confidence on haplotype calls. As an example, Table S1 (Supplementary Data) provides a simulated data set in which each row represents a short read and its called bases, and a "$-$" sign represents a missing base. Applying our proposed model with $e_{ir} = 0.001$ for all reads and bases, we infer that three local haplotypes, AA, GA and GG, are present in the sample using an FDR threshold $f_0 = 0.01$. If we increase the e_{ir} to 0.2 we obtain only one local haplotype AA with $f_0 = 0.01$. If we use $e_{ir} = 0.14$, we get two significant local haplotypes AA and GG.

In LocHap, we remove reads having a mapping quality score less than 30; we also consider a base missing if the Phred quality score of base calling is less than 30. These two steps ensure the high quality of the reads and bases used in the statistical inference. Then we take a conservative value of 0.001 for all the e_{ir}'s as the default setting. This is a conservative choice since 0.001 is the largest possible e_{ir} value after the above read filtering. As a less conservative choice, one could use the provided e_{ir} for each base and read from the *bam* file.

Efficient Computational Algorithm

Posterior inference of LHVs centers at the calculation of $Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs})$. We re-list Equation (1) again to facilitate the subsequent discussion, given by

$$Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}) \propto \sum_{\boldsymbol{\lambda}_{-j} \in \mathcal{Y}_{L-1}} \left[\underbrace{\prod_{i=1}^{N} \underbrace{Pr(\boldsymbol{s}_i^{obs} \mid \lambda_j = 1, \boldsymbol{\lambda}_{-j})}_{\mathcal{I}} \underbrace{Pr(\lambda_j = 1, \boldsymbol{\lambda}_{-j})}_{\mathcal{II}}}_{\mathcal{III}} \right] \tag{4}$$

As mentioned before, if the number of SNVs is R, then the number of possible haplotypes $L = 2^R$, assuming up to two alleles can be observed at each SNV. Correspondingly, we have L number of λ_j's to estimate and the total different configurations of all the λ_j's is $2^L = 2^{2^R}$, a super exponent of R. Therefore, when R is slightly increased, say from 2 to 4, the number of configurations to be calculated increases from 64 to $65,536$. This super-exponential increment calls for efficient computation.

A straightforward way to calculate the right hand side of the Equation (4) would follow the derivation in the previous section, resulting in computing multiple loops of summations and products. It would be time consuming. We take a more efficient approach. For each $j = 1, 2, \ldots, L$, summing over all the binary configurations of $\boldsymbol{\lambda}_{-j}$ amounts to 2^{L-1} many sums. Each term under the outer sum is denoted by $\mathcal{III}$ in (4). A straightforward computation of (4) would calculate term $\mathcal{III}$ $(L * 2^{L-1})$ times for all the j. Same amount of computation is also required for calculation of $Pr(\lambda_j = 0 \mid \boldsymbol{s}^{obs})$ for all $j = 1, 2, \ldots, L$. But careful examination of the terms to be added reveals that some terms are repeatedly calculated L times. For example, assume $L = 4$. In calculating the probability for the event $Pr(\lambda_1 = 1 \mid \boldsymbol{s}^{obs})$, we have to sum over all the other $2^{L-1} = 8$ configurations of $\boldsymbol{\lambda}_{-1}$. Let us take one specific configuration from that set of 8 configurations, $\boldsymbol{\lambda}_{-1} = 101$ (meaning the three elements in $\boldsymbol{\lambda}_{-1}$ take values 1, 0, and 1, respectively). When $\lambda_1 = 1$, the full vector $\boldsymbol{\lambda}$ takes 1101. However, the value 1101 will also show up in the computation of $Pr(\lambda_2 = 1 \mid \boldsymbol{s}^{obs})$ with $\boldsymbol{\lambda}_{-2} = 101$, $Pr(\lambda_3 = 0 \mid \boldsymbol{s}^{obs})$ with $\boldsymbol{\lambda}_{-3} = 111$ and $Pr(\lambda_4 = 1 \mid \boldsymbol{s}^{obs})$ with $\boldsymbol{\lambda}_{-4} = 110$. Therefore, we only need to compute the joint probability of $\boldsymbol{\lambda} = \{1101\}$ once and re-use it for the other three terms. Similarly, for all other possible configurations of $\boldsymbol{\lambda}$, we only need to compute it once. The straightforward way of computation would calculate each configuration four times.

Once all 2^L configurations are calculated, we add up the probabilities from appropriate configurations in order to calculate the probability $Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs})$. We first put decimal indices against all the configurations of $\boldsymbol{\lambda}$ from 0 to $(2^L - 1)$ by treating the 1-st position as the most significant bit (MSB) of a binary string and convert the binary string to its decimal equivalent number. For example, the decimal index of $\boldsymbol{\lambda} = \{\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0, \lambda_4 = 1\}$ is 13. Denote each configuration by C_l where $l = 0, 1, \ldots 2^L - 1$. Once indexing is done, then for each event we sum up the probabilities for a fixed (computed beforehand) set of indices of configurations. For example, for the computation of $Pr(\lambda_2 = 1 \mid \boldsymbol{s}^{obs})$ the set of indices is $\{4, 5, 6, 7, 12, 13, 14, 15\}$. Similarly for $Pr(\lambda_3 = 0 \mid \boldsymbol{s}^{obs})$ that set is $\{0, 1, 4, 5, 8, 9, 12, 13\}$. Denote the set of indices for computing $Pr(\lambda_1 = 1 \mid \boldsymbol{s}^{obs})$ and $Pr(\lambda_1 = 0 \mid \boldsymbol{s}^{obs})$ by $\mathcal{V}_{j1}$ and $\mathcal{V}_{j0}$, respectively. Below, we propose Algorithm 1 for computing (4).

Algorithm 1 Algorithm for computing $Pr\{\lambda_j = 1 \mid \boldsymbol{s}^{obs}\} \quad \forall j = 1, 2, \ldots, L$

```
Index all the configurations of $\boldsymbol{\lambda}$ from 0 to $(2^L - 1)$.
Enumerate $\mathcal{V}_{j1}$ and $\mathcal{V}_{j0}$ for all $j = 1, 2, \ldots, L$.
Compute $Pr(\boldsymbol{s}^{obs} \mid C_l)$ and $Pr(C_l)$ for all $l = 0, 1, \ldots, 2^L - 1$.
for $j = 1, 2, \ldots, L$ do
    $P_1(j) \leftarrow \sum_{l \in \mathcal{V}_{j1}} Pr(\boldsymbol{s}^{obs} \mid C_l)\ Pr(C_l)$.
    $P_0(j) \leftarrow \sum_{l \in \mathcal{V}_{j0}} Pr(\boldsymbol{s}^{obs} \mid C_l)\ Pr(C_l)$.
    $Pr(\lambda_j = 1 \mid \boldsymbol{s}^{obs}) \leftarrow \frac{P_1(j)}{P_1(j) + P_0(j)}$.
end for
```

Calculation of $Pr(\boldsymbol{s}^{obs} \mid C_l)$ for all $l = 0, 1, \ldots, 2^L - 1$ in Algorithm 1 is carried out with an additional algorithm that takes advantage of the structured probability formulation. The detail is shown in Supplementary Data. Calculation of $Pr(C_l)$ is trivial based on the independent prior of λ_j's. Because of the closed-form derivation for (4) and efficient computation algorithms, LocHap is ultra-fast in analyzing whole-genome and whole-exome data, taking usually less than a minute for a whole genome.

LocHap pipeline

A computational pipeline (Fig. 3) supports LocHap applications. LocHap analyzes one sample at a time and can be used sequentially or in parallel for the analysis of multiple samples. For a single sample, the input of LocHap includes 1) a *bam* file with the associated index *bai* file and 2) a corresponding *vcf* file that contains the SNVs in the sample, called by any standard variant calling algorithm, such as *GATK's* [27–29] UnifiedGenotyper tool. The output of LocHap is a set of LHVs stored in *hcf* files, one for each sample.

Haplotype Call Format (hcf)

Each line in the *hcf* file contains information about one particular LHV segment. Below is a line in an *hcf* file from analysis of real-world data.

```
##CHROM   POS   REF   NumSig   HAP_Call   All_HAP   DataForSample = NA12878
chr1   4369613, 4369623   GA   3   GA(1.000), AA(1.000), GG(0.985)   GA(1.000; 0.000), AA(1.000; 0.000), GG(0.985; 0.005), AG(0.000; 0.254)
nSNP = 2; nTot = 90; nACGT = 75; nBlank = 15; nDisc = 0; nM0 = 41; nM1 = 34; nM2 = 15; nClus = 3;
```

Same as *vcf*, an *hcf* file is a tab-delimited text file. After the initial header fields, each line in the hcf file represents a local haplotype (might not be a variant) and has seven column fields. Also, at the end of each *hcf* file, a summary stating the total number of SNVs in the *vcf* file, number of segments with zero significant haplotypes, one significant haplotype, two significant haplotypes and so on, are given (see Supplemental Data).

Post Processing

The inferred LHVs can be filtered to remove false positives caused by artifact and noise in the sequencing data. The filters are devised to remove dubious reads and SNVs. In NGS data, known artifact and error affect SNV calling [21] by erroneously calling or aligning the bases on short reads. However, they do not artificially create additional local haplotypes which require more sophisticated changes to the bases of short reads across multiple SNVs. The typical artifact and error often changes the base calling or alignment for a single locus and usually affect all the short reads in the region. Therefore, LHV-based inference is less prone to the errors and artifacts for SNV calling. Nonetheless, we apply a set of optional and customizable filters (Supplementary Data) with different stringency levels for post-processing of the LHVs in the output *hcf* files. Currently, despite the large amount of effort directed by the community [21] the noise and error in NGS experiments and data preprocessing cannot be statistically modeled or quantified. There is no consensus on filtering the variant calls from various analysis pipelines. We present a conservative filtering pipeline that is heavily biased towards reducing FDR, so that reported LHVs are of high confidence. The proposed filtering depends on various parameters that can be modified to enforce different degrees of filter stringency. A more stringent filter results in fewer LHVs at the end.

The proposed filters can remove SNVs that are too close to each other (within, say 50 bps) and SNVs that are close to other types of variants such as indels. It has been noted [30, 31] that these variant calls are not trustworthy due to artifacts and base calling errors in the data. In addition, our filters can remove SNVs for which most reads are aligned to the SNV at a base near the end of the reads. The reason is that bases called towards the end of a read are usually of low quality, which then affect the reliability of the alignment. Lastly, SNVs mapped by reads with strand bias [32] are also filtered.

Integrated Genome Viewer (IGV) Compatibility

For better visualization, we provide an additional IGV-compatible format so that LHV segments can be visually examined in the popular genome visualization tool IGV [22]. A snapshot of five *hcf* files in IGV is shown in Fig. 4. The details of the corresponding command is given in *Quick Manual* (`http://compgenome.org/lochap/code_release/QuickManual-LocHap-release-v1.0.pdf`). Note that the LHVs are shown by *red* bars and non variants are shown by *blue* bars.

RESULTS AND DISCUSSION

Simulation

We first demonstrate the utility of the proposed statistical model using simulated data. In all of the following examples we used $e_{ir} = 0.001$ as the default value and FDR threshold was set at $f_0 = 0.01$. Also we assumed that the probability of observing more than two different alleles at a particular locus in a genome was considered rare. This is assumed because of the small chance of having a point mutation occurring twice at the same nucleotide [33–35]. All of the simulation examples were based on short reads data generated for a single segment. Also, we only show examples with a small number of short reads. When a large number of reads were simulated, the proposed models performed very well, easily recovering the simulation truth. Tabulated posterior probabilities for all the scenarios are provided in the Supplementary Data (Tables S1-S9).

Simulation Scenario 1 We generated eight short reads covering two SNVs. Assuming at each SNV that only two alleles could be observed, there were $2^2 = 4$ possible haplotypes. The simulated short reads had genotypes

$$D = \{GA, GA, GA, GA, GC, GC, AC, AC\}.$$

Applying LocHap, we inferred three significant haplotypes with the following sequences and posterior probabilities.

$$\{GA : 1.00,\ GC : > 0.99,\ AC : > 0.99\}$$

Simulation Scenario 2 In this scenario, we generated eight short reads, each of which only covered one of the two SNVs.

$$D = \{A-, -A, G-, G-, -G, -G, -G, -G\}.$$

Here the read labeled "$A-$" indicated that the first SNV position had a readout A and the read did not cover the second SNV position. Hence, we used "$-$" sign to represent a missing genotype. Using LocHap, no haplotypes can be inferred to be present based on the FDR threshold $f_0 = 0.01$.

Simulation Scenario 3 In this scenario, we simulated five short reads covering three SNVs with genotypes given by

$$D = \{AGA, AGA, AGC, AGC, GAC\}.$$

LocHap called three significant haplotypes $\{AGA, AGC, GAC\}$ with posterior probabilities all > 0.99.

Simulation Scenario 4 In this scenario, we generated eight short reads covering three SNVs, given by

$$D = \{AAA, AAT, ACA, ACT, GAA, GAT, GCA, GCT\}.$$

LocHap did not identify any significant haplotypes. This is due to the lack of strong evidence for any of the haplotypes as each of them is supported by only one read. The proposed model correctly recognized the uncertainty in the data and did not provide statistical significance for any haplotypes.

Simulation Scenario 5 This scenario is similar to scenario 4 but here we generated five short reads each for the haplotypes AAA, AAT and ACA. The data is given by

$$D = \{AAA \times 5, AAT \times 5, ACA \times 5, ACT, GAA, GAT, GCA, GCT\}.$$

Although LocHap did not find any significant haplotypes in the previous scenario, LocHap called three significant haplotypes $\{AAA, AAT, ACA\}$ with posterior probabilities all equal to 1.00 because of more reads are generated for these three haplotypes. It is easy to see that with more number of reads our inference model would work more accurately.

Simulation Scenario 6 This scenario is similar to scenario 1 but here we generated five times more number of short reads for every categories. The data is given by

$$D = \{GA \times 20, GC \times 10, AC \times 10\}.$$

Applying LocHap, we again inferred three significant haplotypes with posterior probabilities all equal to 1.00.

Simulation Scenario 7 This scenario is same as the real-life data in Fig.1(a). The data is given by

$$D = \{GG \times 23, GC \times 10, AG \times 9, -G \times 43, G - \times 26, A - \times 13, -C \times 11\}.$$

Applying LocHap, we inferred three significant haplotypes $\{GG, GC, AG\}$with posterior probabilities all equal to1.00.

Simulation Scenario 8 This scenario is same as the real-life data Fig.1(b). The data is given by

$$D = \{TT \times 22, CG \times 14, CT \times 13, -T \times 46, -G \times 46, C - \times 34, T - \times 12\}.$$

Applying LocHap, we again inferred three significant haplotypes $\{TT, CG, CT\}$ with posterior probabilities all equal to 1.00.
All eight scenarios show that LocHap performs well. When the number of reads increases, the confidence in the statistical inference also increases.

Three DNA-Seq Data Sets

We applied LocHap to three different data sets, among which two were public and one from our own in-house validation experiments. We provide main findings next and put analysis details in the Supplementary Data.

Head and Neck Cancer (HNC) Data We analyzed WES data of 30 matched tumor and blood sample pairs (total 60 samples) from patients with head and neck cancer [36]. Whole exome Sequence Read Archive (SRA) files of matched tumor and normal samples were downloaded from the Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`). Standard bioinformatics analyses were performed to extract *fast-q* sequences, map short reads, and call SNVs. We generated *bam* files (one per sample) and a *vcf* file for all the samples. The *bam* files contained short read sequences and alignments, and the *vcf* file contained SNV calls of all the samples. The *bam* files with associated *bai* and *vcf* files were provided to the LocHap pipeline, which subsequently generated 60 *hcf* files, one for each sample.

Fig. 5a shows a circos plot [37] of the called LHVs. Most LHVs are located in different genomic regions across patients, suggesting somatic mutations occurred randomly across the genome. Also, the fact that called LHVs are mostly different between patients indirectly shows that LHV calling is not driven by artifact and noise in the NGS data. The reported LHVs all passed the aforementioned noise filtering with stringent criteria. Read depth of one exome of a normal sample is shown in Fig. 5b. A few LHVs are mapped with large numbers of reads but overall the read depths between LHVs and non-variant regions are comparable. Most LHs are not LHVs, having no more than two genotypes and most LHVs possess three genotypes (Fig. 5c). Tumors in general possess more LHVs than corresponding normal samples (Fig. 5d) and chromosomes 9, 14 and 17 are "hotspots" for LHVs exhibiting higher frequencies in tumors than blood samples (Fig. 5e). Transitions are more frequent than transversions (Fig. 5f), as expected. Finally, overlapping LHVs are present in both tumor and the matched blood samples for each of the 30 patients (Fig. 6), while the tumor and blood samples also possess unique LHVs of their own.

Table 1 summarizes the statistics from the unfiltered *hcf* file from one particular normal blood sample.

CEU-TRIO Data from 1,000-Genomes Project We applied LocHap to WGS data of a CEU TRIO family of father, mother and child from the 1,000-Genomes project (`http://www.1000genomes.org/`). The analysis procedure was identical to the HNC data, except here we have WGS data from three members of a family. Genome-wide LHVs (Fig. 7a) are found in all three individuals with father having the largest number of LHVs and daughter the smallest (Fig. 7b). This reflects the evolutionary conjecture that somatic mutations emerge over time as a result of accumulating mitotic errors and that the longer an individual lives, the more likely somatic mosaicism is seen on the genome [18]. Similar to the results obtained in the previous analysis of cancer WES data, most LHs are not LHVs and most LHVs possess three genotypes (Fig. 7c). Most LHVs reside in intergenic and intronic regions with less than 1% in exons (Fig. 7d). Here again transitions are more prevalent than transversions (Fig. 7e). We called CNVs using CNVnator [38]. Convincingly, CNVs are not observed for most LHVs regions, suggesting that the LHVs are not associated with CNVs, a potential confounder for LHV calling. There are almost no overlapping LHVs across the three family members. This is expected since LHVs are results of somatic mutations, which do not usually re-occur in different individuals. Under the most stringent filter, on average 400-500 LHVs are reported per genome using the WGS data in CEU trio compared with 4-5 per exome using the WES data from the HNC sample.

Table 2 summarizes the statistics from the unfiltered *hcf* file from one particular sample (NA12891) in this dataset.

Validation In order to validate our results, we sequenced whole blood DNA from three members of a parent-child trio using two different sequencing platforms, Complete Genomics, Inc. (CGI) (`http://www.completegenomics.com/documents/DataFileFormats_Standard_Pipeline_2.5.pdf`, `http://cgatools.sourceforge.net/docs/1.8.0/cgatools-user-guide.pdf`) and Illumina whole genome sequencing (ILMN) (`http://www.illumina.com/applications/sequencing.html`). All members of the trio were healthy. Their blood samples were collected between 2007 and 2012 and sequenced by CGI in 2012 [39]. We also sequenced DNA from the same three samples using the ILMN platform in 2014.

Because ILMN and CGI utilized different sequencing technologies and the sequencing experiments were performed at separate times by more than two years apart, results from the two sequencing experiments serve to validate each other.

NGS data produced by both technologies were analyzed using the LocHap pipeline. At the end, for each of the two data sets we generated a list of LHV. We then overlapped the two lists of LHVs, and identified shared LHVs between both data sets. Nine LHVs overlapped between the two data sets in the child; 10 LHVs overlapped in the mother and 15 LHVs overlapped in the father. We applied highly stringent filtering rules (Supplemental Data) to ensure high quality of the reported LHVs, although such filtering could also remove true LHVs with weak confidence. Also, many LHVs were excluded due to insufficient evidence from the CGI data. Fig. 8 shows the locations of LHVs for the CGI and ILMN data. Fig. 9 presents two LHV examples that are shared between CGI and ILMN data. For these two LHVs, the short reads provide direct evidence of somatic mosaicism – the reads suggest that at least three local haplotypes must be present.

These analyses provide evidence supporting our hypothesis that normal cells in a healthy person could be genetically heterogeneous and possess distinct populations of somatic cells, a phenomenon also observed in [5]. Specifically, in all three individuals there are LHVs that are discovered by independent sequencing platforms from different experimentalists at different times.

Discussion

Through a novel means of analyzing NGS data, LocHap attempts to reveal potential somatic mosaicism in the form of LHVs. We implement Bayesian hierarchical models that borrow strength from the mapped short reads to infer the number and sequences of LHVs genome-wide. In applications of LocHap using deep-sequencing data, we provide evidence that supports the existence of normal somatic mosaicism (NSM) and tumor somatic mosaicism (TSM) at single-nucleotide level. Applying LocHap to 30 matched blood and tumor samples, we find LHVs in exomes of normal blood and tumor samples. The frequencies of LHVs are in general higher in tumor samples (one-sided paired t-test, p-value < 0.0001). Performing the analysis on CEU trio from the 1,000-genome project, we confirm the findings of genome-wide LHVs and also identify an increasing trend of LHV occurrences with aging (chi-squared test [40] for trend, p-value < 0.0001). Based on our results, we propose three hypotheses that deserve future investigation.

1. Similar to cancer cells, non-cancer cells undergo random mutation events that could potentially lead to subclonal cell populations, resulting in genome-wide somatic mosaicism within individuals.

2. The probability of acquiring NSM increases with the age of an individual, owning to accumulating mutation burden.

3. In general, TSM is more prevalent than NSM.

LocHap is different from existing subclonal callers [35,41–47] in a fundamentally distinct way. LocHap provides direct evidence (e.g., examples in Fig. 1) of genome mosaicism in both non-cancer and cancer cells. The units of analysis under LocHap are haplotypes, each as a scaffold of SNVs. In contrast, most subclone callers in the literature analyze allelic fractions of individual SNVs. We argue that LocHap provides a more direct view on genome mosaicism for somatic samples. The power of detecting LHVs is affected by the length of paired-end reads and coverage. In this context, it is important to note that an unexpected insert size in a paired-end read is handled by either initial choice of the value K (if it is too large) or by using a post-processing filter (if it is too small).In addition, our analysis does not include paired-end reads that are not properly mapped since these reads do not provide reliable information about LHVs.

Naturally, sequencing coverage, quality and read length affect the performance of LocHap. Deeper coverage allows local haplotype variants with small population frequencies in the sample to be detected. Longer read length (and/or insert length) allows SNVs that are farther apart to be phased and therefore improve the chance of detecting LHVs. In our WES data our coverage was about 30X with read length 75bps and in WGS data we have about 60X coverage with read length 100bps. We found that our LocHap performed well in reporting LHVs under these conditions. For detail of all the bioinformatics pipelines, filters and parameters, we refer the readers to the Supplemental data.

Our main purpose is not to identify all the LHVs in the genome. Instead, we aim to utilize existing short-read NGS data and provide a new method for detecting sample heterogeneity and mosaicism based on LHVs. Presence of LHVs itself supports mosaicism since a homogeneous human biological sample cannot harbor LHVs.

LocHap is available at `http://www.compgenome.org/lochap/` for free download. A manual is provided along with the software. It is ultrafast in calling LHVs. For one WES sample with about 30X depth of coverage, whole-exome LHV calling by LocHap took about 11 seconds on a Macbook Pro (2.8GHZ Intel Core i7 and 16GB 1600 MHz DDR3 memory). For each WGS data set with about 60X depth of coverage the analysis took about 47 seconds.

Cellular mosaicism based on LHVs would facilitate studies on heterogeneity of cell populations. Availability of NGS data allows for more powerful investigation of somatic cell subpopulations. The resolution of analysis can be at single nucleotide level, as opposed to mega-bases for microarray data. Further validation of somatic mosaicism and its relationship to aging and diseases is needed using much bigger sample sizes. Such effort could help us reveal and quantify heterogeneity in non-cancer and cancer samples, potentially affecting cancer diagnosis and prognosis.

CONCLUSION

Through LocHap we provide a new approach to extract information of local haplotypes from NGS data for a single sample. We found wide-spread LHVs across genome in both tumor and non-tumor samples. These results and software tools can be used for further investigation of somatic mosaicism in human samples, helping investigators to understand the frequency and genome locations of mosaic events. Thanks to the ultrafast speed of LocHap, it can be used to analyze a large number of samples using a single computer or a small cluster. The newly developed *hcf* files follow the existing format standards for *vcf* files and can be visualized in the popular tool IGV.

References

[1] M Gerlinger, AJ Rowan, S Horswell, J Larkin, D Endesfelder, E Gronroos, P Martinez, N Matthews, A Stewart, P Tarpey, I Varela, B Phillimore, S Begum, NQ McDonald, A Butler, D Jones, K Raine, C Latimer, CR Santos, M Nohadani, AC Eklund, B Spencer-Dene, G Clark, L Pickering, G Stamp, M Gore, Z Szallasi, J Downward, PA Futreal, and C Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.*, 366(10):883–892, 2012.

[2] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[3] James R Lupski. Genome Mosaicism–One Human, Multiple Genomes. *Science*, 341(6144):358–359, 2013.

[4] Hagop Youssoufian and Reed E Pyeritz. Mechanisms and consequences of somatic mosaicism in humans. *Nature Reviews Genetics*, 3(10):748–758, 2002.

[5] Giulio Genovese, Anna K Kähler, Robert E Handsaker, Johan Lindberg, Samuel A Rose, Samuel F Bakhoum, Kimberly Chambert, Eran Mick, Benjamin M Neale, Menachem Fromer, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26):2477–2487, 2014.

[6] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[7] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[8] Hege G. Russnes, N Navin, J Hicks, and Anne-Lise Borresen-Dale. Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of Clinical Investigation*, 121(10):3810–3818, 2011.

[9] Nicholas Navin, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, Yvonne Eberling, Jennifer Troge, Vladimir Grubor, et al. Inferring tumor progression from genomic heterogeneity. *Genome research*, 20(1):68–80, 2010.

[10] Steven A Frank. Somatic mosaicism and cancer: inference based on a conditional Luria–Delbrück distribution. *Journal of theoretical biology*, 223(4):405–412, 2003.

[11] Steven A Frank. Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1725–1730, 2010.

[12] Steven A Frank. Somatic Mosaicism and Disease. *Current Biology*, 24(12):R577–R581, 2014.

[13] Maeve O'Huallachain, Konrad J Karczewski, Sherman M Weissman, Alexander Eckehart Urban, and Michael P Snyder. Extensive genetic variation in somatic human tissues. *Proceedings of the National Academy of Sciences*, 109(44):18018–18023, 2012.

[14] Alexej Abyzov, Jessica Mariani, Dean Palejev, Ying Zhang, Michael Seamus Haney, Livia Tomasini, Anthony F Ferrandino, Lior A Rosenberg Belmaker, Anna Szekely, Michael Wilson, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature*, 492:438–442, 2012.

[15] Naoki Nagata and Shinya Yamanaka. Perspectives for Induced Pluripotent Stem Cell Technology New Insights Into Human Physiology Involved in Somatic Mosaicism. *Circulation research*, 114(3):505–510, 2014.

[16] Lars A Forsberg, Chiara Rasi, Hamid R Razzaghian, Geeta Pakalapati, Lindsay Waite, Krista Stanton Thilbeault, Anna Ronowicz, Nathan E Wineinger, Hemant K Tiwari, Dorret Boomsma, et al. Age-related somatic structural changes in the nuclear genome of human blood cells. *The American Journal of Human Genetics*, 90(2):217–228, 2012.

[17] Cathy C Laurie, Cecelia A Laurie, Kenneth Rice, Kimberly F Doheny, Leila R Zelnick, Caitlin P McHugh, Hua Ling, Kurt N Hetrick, Elizabeth W Pugh, Chris Amos, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature genetics*, 44(6):642–650, 2012.

[18] Kevin B Jacobs, Meredith Yeager, Weiyin Zhou, Sholom Wacholder, Zhaoming Wang, Benjamin Rodriguez-Santiago, Amy Hutchinson, Xiang Deng, Chenwei Liu, Marie-Josephe Horner, et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nature genetics*, 44(6):651–658, 2012.

[19] Ursula M Schick, Andrew McDavid, Paul K Crane, Noah Weston, Kelly Ehrlich, Katherine M Newton, Robert Wallace, Ebony Bookman, Tabitha Harrison, Aaron Aragaki, et al. Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PloS one*, 8(3):e59823, 2013.

[20] Lars A Forsberg, Chiara Rasi, Niklas Malmqvist, Hanna Davies, Saichand Pasupulati, Geeta Pakalapati, Johanna Sandgren, Teresita Diaz de Ståhl, Ammar Zaghlool, Vilmantas Giedraitis, et al. Mosaic loss of chromosome y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature genetics*, 46(6):624–628, 2014.

[21] Heng Li. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30:2843–2851, 2014.

[22] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[23] Michael A Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[24] Peter Müller, Giovanni Parmigiani, Christian Robert, and Judith Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001, 2004.

[25] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[26] Yuan Ji, Yanxun Xu, Qiong Zhang, Kam-Wah Tsui, Yuan Yuan, Clift Norris Jr, Shoudan Liang, and Han Liang. BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data. *Biometrics*, 67(4):1215–1224, 2011.

[27] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[28] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[29] Geraldine A Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, pages 11–10, 2013.

[30] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, 12(1):451, 2011.

[31] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[32] Yan Guo, Jiang Li, Chung-I Li, Jirong Long, David C Samuels, and Yu Shyr. The effect of strand bias in Illumina short-read sequencing data. *BMC genomics*, 13(1):666, 2012.

[33] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[34] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[35] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):35, 2014.

[36] Nicolas Stransky, Ann Marie Egloff, Aaron D Tward, Aleksandar D Kostic, Kristian Cibulskis, Andrey Sivachenko, Gregory V Kryukov, Michael S Lawrence, Carrie Sougnez, Aaron McKenna, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–1160, 2011.

[37] Martin Krzywinski, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[38] Alexej Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–984, 2011.

[39] Oren E Livne, Lide Han, Gorka Alkorta-Aranburu, William Wentworth-Sheilds, Mark Abney, Carole Ober, and Dan L Nicolae. Primal: Fast and accurate pedigree-based imputation from sequence data in a founder population. *PLoS computational biology*, 11(3):e1004139–e1004139, 2015.

[40] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2013.

[41] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[42] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol*, 14(7):R80, 2013.

[43] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17):e165–e165, 2013.

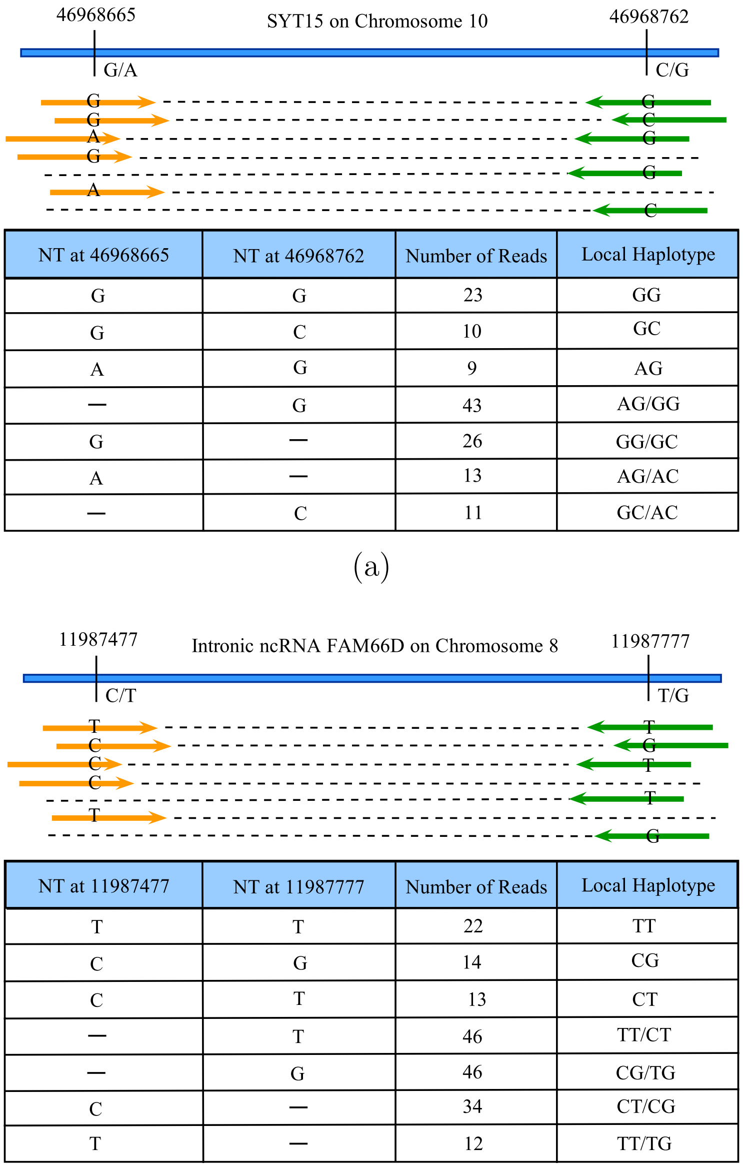
[44] Mengjie Chen, Murat Gunel, and Hongyu Zhao. Somatica: Identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. *PloS one*, 8(11):e78143, 2013.

[45] Noemi Andor, Julie V Harness, Sabine Mueller, Hans W Mewes, and Claudia Petritsch. Expands: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, 2014.

[46] Andrej Fischer, Ignacio Vázquez-García, Christopher JR Illingworth, and Ville Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell reports*, 7(5):1740–1752, 2014.

[47] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun H Jang, Lincoln Stein, and Quaid Morris. Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015.

Figures



NT at 46968665	NT at 46968762	Number of Reads	Local Haplotype
G	G	23	GG
G	C	10	GC
A	G	9	AG
—	G	43	AG/GG
G	—	26	GG/GC
A	—	13	AG/AC
—	C	11	GC/AC

(a)

NT at 11987477	NT at 11987777	Number of Reads	Local Haplotype
T	T	22	TT
C	G	14	CG
C	T	13	CT
—	T	46	TT/CT
—	G	46	CG/TG
C	—	34	CT/CG
T	—	12	TT/TG

(b)

Figure 1: Two examples of LHVs based on direct observations of aligned short reads. The pairs of a single short read are marked with orange and green colored arrow, respectively. Panel (a): an LHV called from WES data of a normal blood sample. The haplotype consists of two SNVs separated by 97 base pairs in a coding region of gene *SYT15*. Among all the short reads mapped to this region, 23, 10 and 9 short reads are mapped to both SNVs and exhibit alleles GG, GC, and AG, respectively. Due to the large count of the least frequent allele (AG) and the combined information from all other short reads, LocHap calls three local haplotypes with high statistical confidence, making it a variant (i.e., an LHV). Panel (b): an LHV called from WGS data of a normal blood sample from a normal individual (NA12878 in the CEU TRIO family in the 1,000 genome project). The local haplotype consists of two SNVs separated by 300 base pairs in an intronic region of an ncRNA FAM66D. Again, similar haplotype variants are seen based on the short reads mapped to both SNVs. In both examples, some reads are mapped to only one of the two SNVs. These reads provide partial information on the existence of certain haplotypes. For example, reads with “-G” in panel (a) are only mapped to the second SNV with genotype “G”. They support that haplotypes AG or GG might be present in the sample. Hence, reads mapped to both SNVs and reads mapped to at least one SNV are used in the statistical models of LocHap.

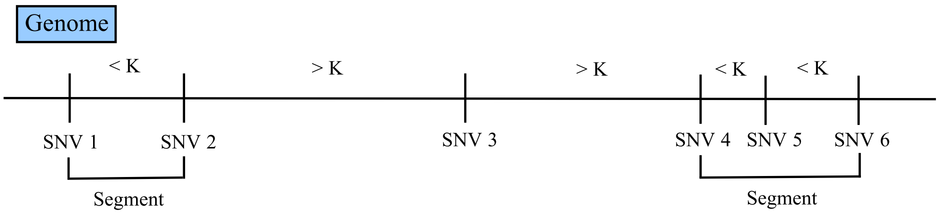


Figure 2: Illustration of DNA segments in LocHap. The first segment consists of two SNVs (SNV 1 and 2) and the second one has three SNVs (SNV 3, 4 and 5). SNV 3 is more than K base pairs from its adjacent SNVs 2 & 5, and therefore is not included in any segment.

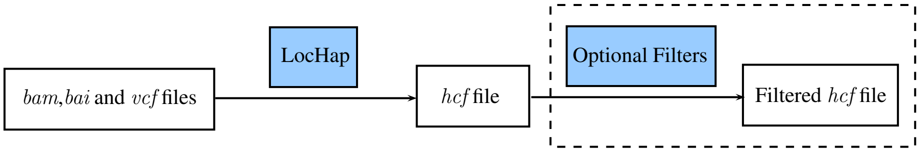


Figure 3: Overview of the LocHap Pipeline.

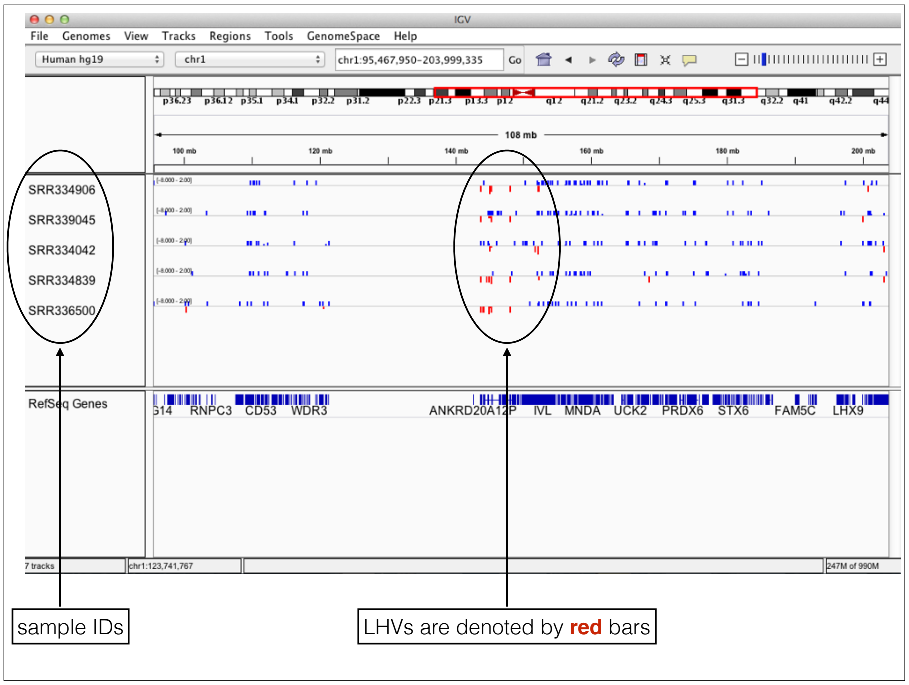
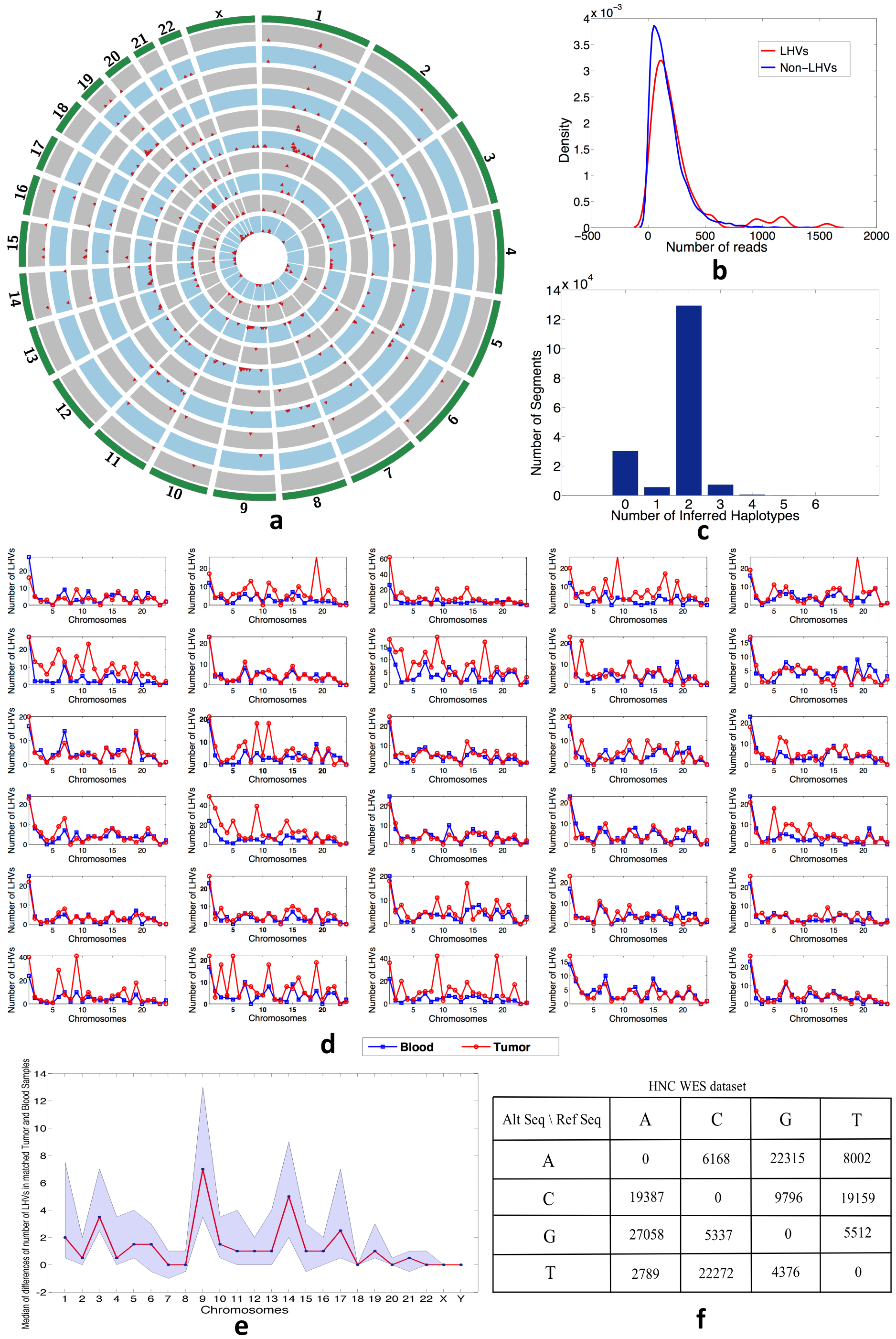

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr1
chr1:95,467,950-203,999,335
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.2 q23.2 q24.3 q25.3 q31.3 q32.2 q41 q42.2 q44
108 mb
100 mb
120 mb
140 mb
160 mb
180 mb
200 mb
SRR334906
SRR339045
SRR334042
SRR334839
SRR336500
RefSeq Genes
514 RNPC3 CD53 WDR3
ANKRD20A12P IVL MNDA UCK2 PRDX6 STX6 FAM5C LHX9
7 tracks
chr1:123,741,767
247M of 990M
sample IDs
LHVs are denoted by **red** bars

Figure 4: Visualization of *hcf* files in IGV.



HNC WES dataset

Alt Seq \ Ref Seq	A	C	G	T
A	0	6168	22315	8002
C	19387	0	9796	19159
G	27058	5337	0	5512
T	2789	22272	4376	0

Figure 5: LHV calling for a head & neck cancer WES data set with 30 pairs of matched tumor and normal samples. (a) A circos plot of prevalence of LHVs for 5 arbitrarily selected sample pairs. Each red dot indicates the existence of at least one LHV in the corresponding exonic region of 1M bps. The height of a red dot indicates the number of LHVs present in the segment of 1M bps long. A pair of matched tumor and normal samples are arranged as adjacent circles with grey and blue color, respectively. (b) Comparison of read depth for genome regions with and without LHVs. No apparent difference is observed. (c) Histogram showing the frequencies of DNA segments (vertical axis) with different numbers of haplotype calls (horizontal axis). Most regions have up to two haplotypes, i.e., no variants. Regions with greater than two haplotypes are variants implying genome mosaicism. (d) A total of 30 line plots, one for each pair of matched tumor (red) and normal (blue) samples from an individual patient. The number of LHVs is shown for each chromosome for each patient. In general, tumors exhibit more LHVs implying more mosaicism. (e) Summary of (d). For each chromosome, a blue dot is the median of the difference in the number of LHVs between tumor and its matched normal sample across 30 patients; point-wise confidence intervals are shown as purple bands. Tumors show much higher frequencies of LHVs on chromosomes 9, 14, and 17, indicating potential disease-related variations on these regions. (f) Summary of sequence mutations for the SNVs within called LHVs. Transitions are much more prevalent than transversions.

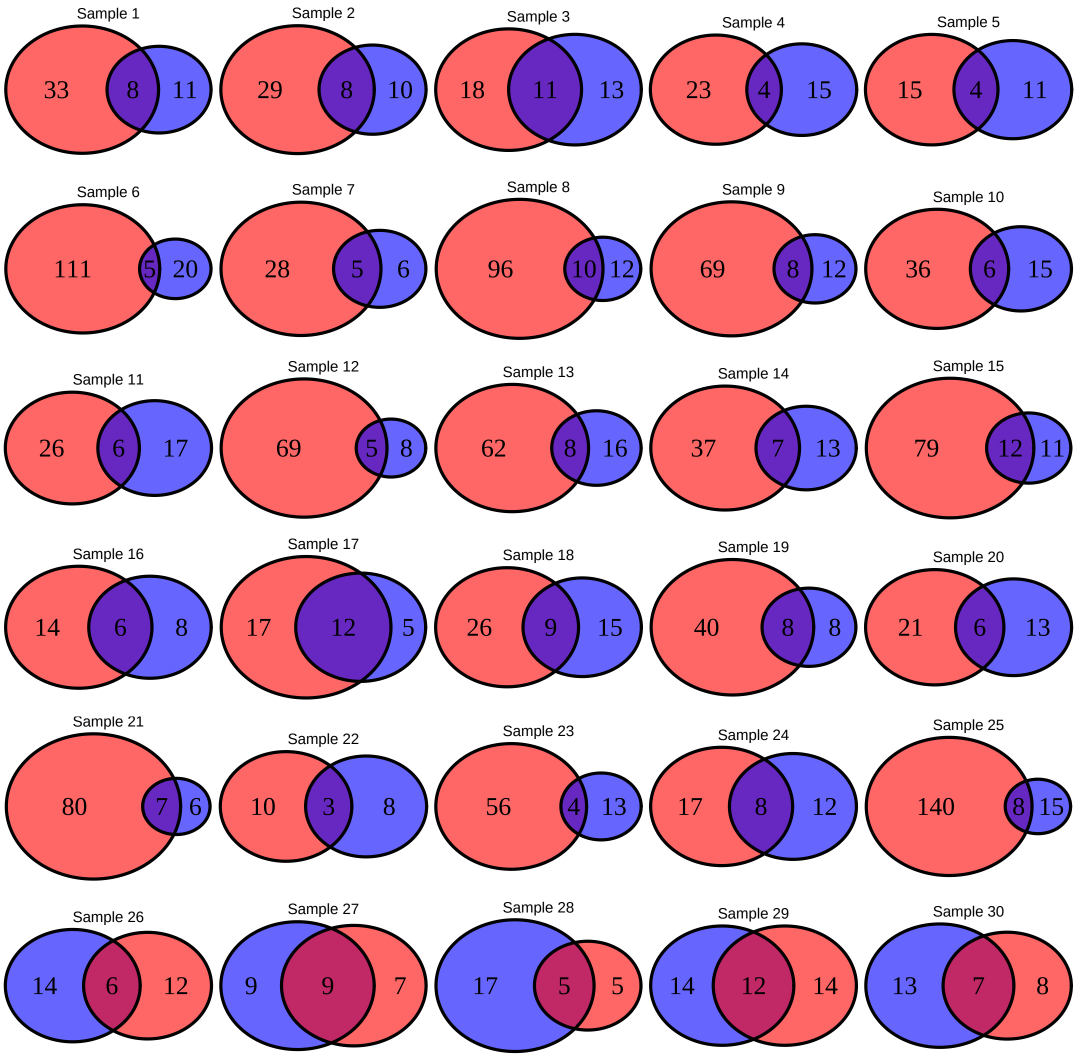
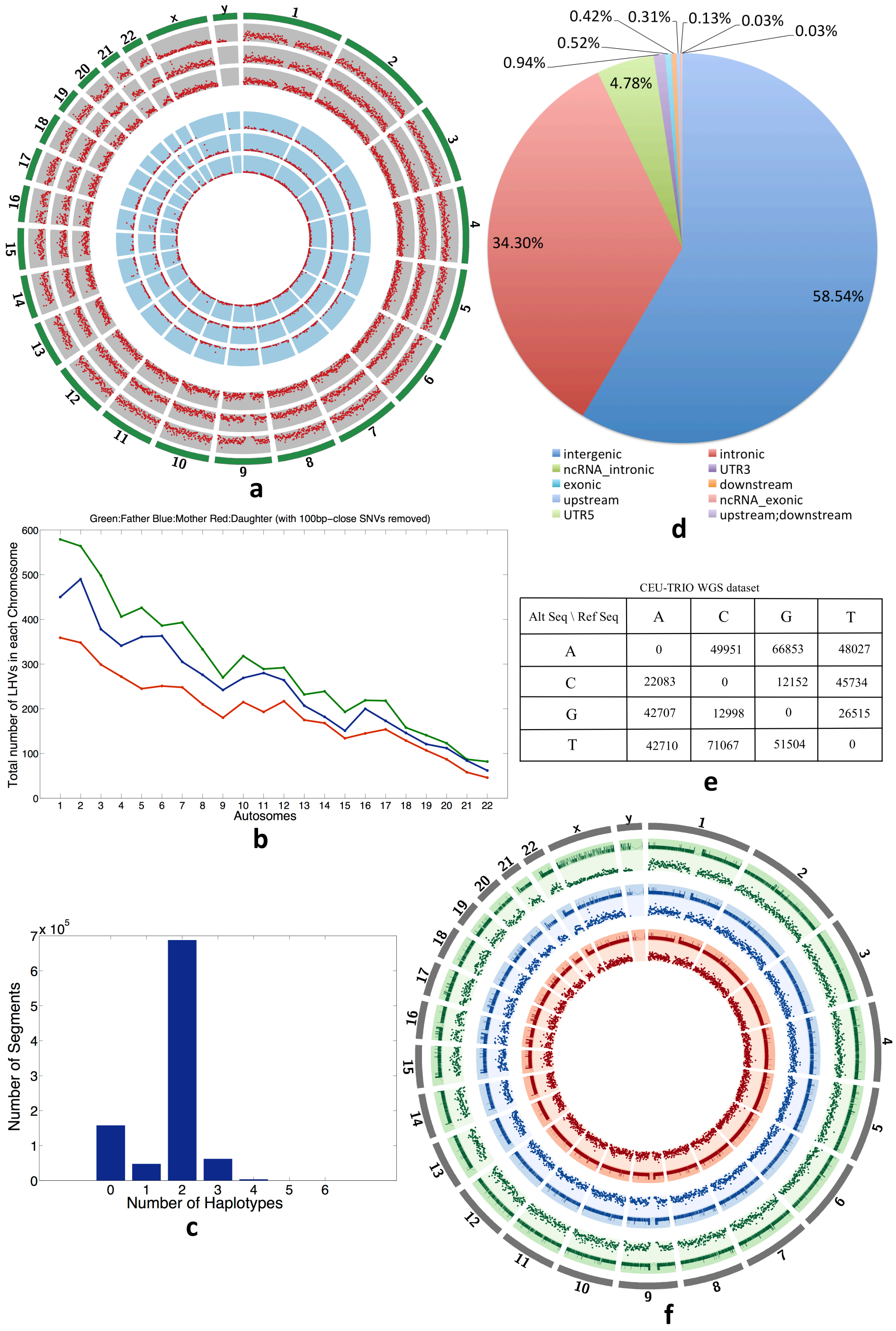


Figure 6: Venn diagrams showing the overlap of LHV calling for a head & neck cancer WES data set with 30 pairs of matched tumor and normal samples. For each pair, LHV counts for tumor and the matched normal sample are shown in red and blue color, respectively. In most of the samples, number of LHVs in tumor is greater than that of the matched normal except for the last 5 samples where the numbers are comparable or number of LHVs found in the tumor sample is less the corresponding number in the matched normal sample.



CEU-TRIO WGS dataset

Alt Seq \ Ref Seq	A	C	G	T
A	0	49951	66853	48027
C	22083	0	12152	45734
G	42707	12998	0	26515
T	42710	71067	51504	0

Figure 7: LHV calls for normal samples from a CEU trio of father, mother, and daughter in the 1,000 genome project based on WGS data. (a) A circos plot of prevalence of LHVs. Outer 3 arcs and inner 3 arcs represent results of TRIO samples filtered by type III filter and type I filter, respectively. See Materials and Methods section for details of the filters. Each red dot indicates the existence of at least one LHV in the corresponding genomic region of 1M bps. The height of a red dot indicates the number of LHVs present in the region. (b) Comparison of the three family members in the number of LHVs per chromosome. The daughter has the smallest and the father has the largest number of LHVs in all chromosomes (autosome). (c) Histogram showing the frequencies of DNA segments (vertical axis) with different numbers of haplotype calls (horizontal axis). Most segments have up to two haplotypes indicating no variant. Segments with greater than two haplotypes are variants implying genome mosaicism. (d) Functional annotations of the genome regions where LHVs are found. Most are intergenic and intronic, with $< 1\%$ LHVs in exons. (e) Summary of sequence mutations for the SNVs within called LHVs. Transitions are much more prevalent than transversions. (f) Copy number calls based on CNVnator [38] are directly compared with LHVs for all three family members. In most cases, there are no copy number variations on genome regions where LHVs are found. Copy numbers are represented in the outer arc and LHVs are shown in the adjacent inner arc in the same color for each sample.

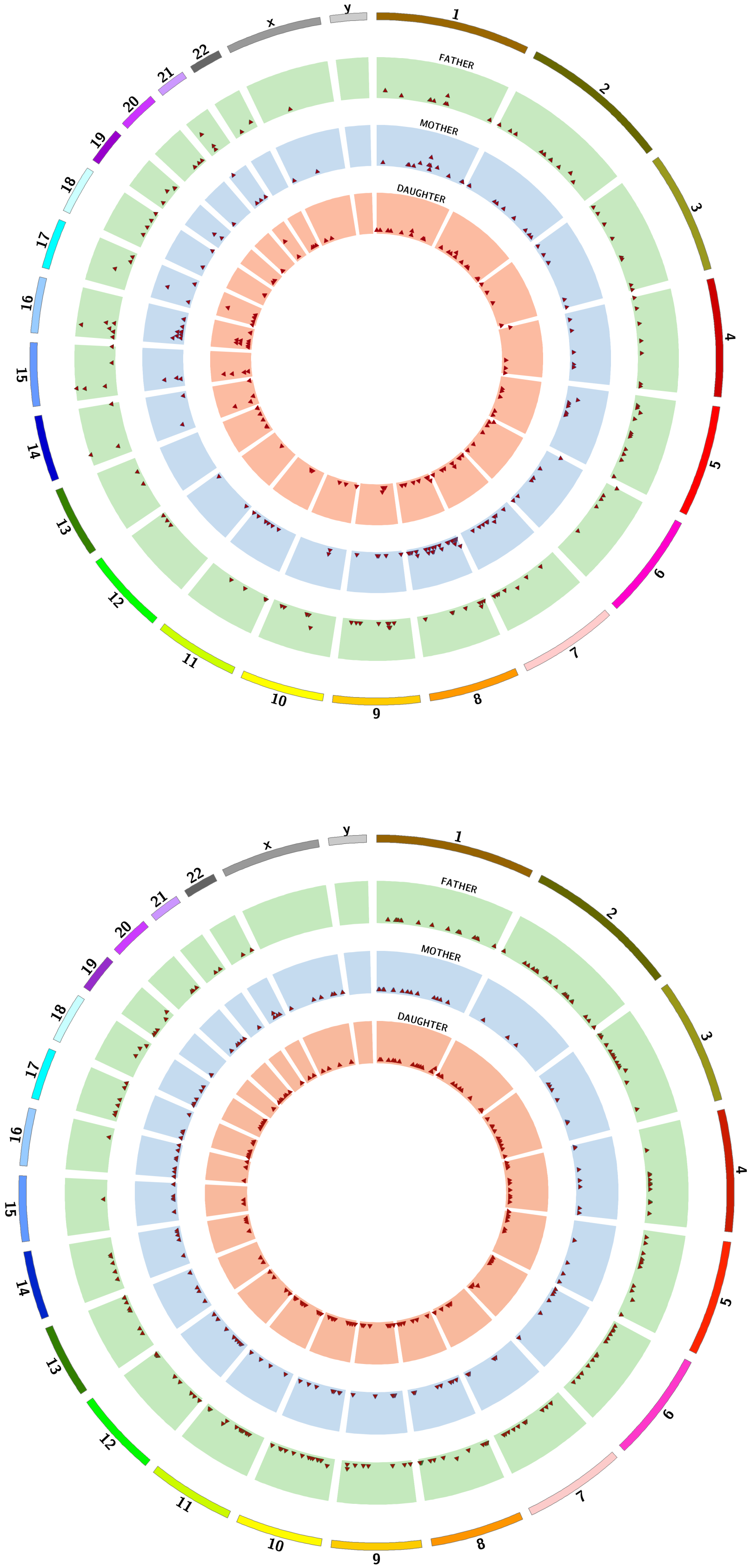
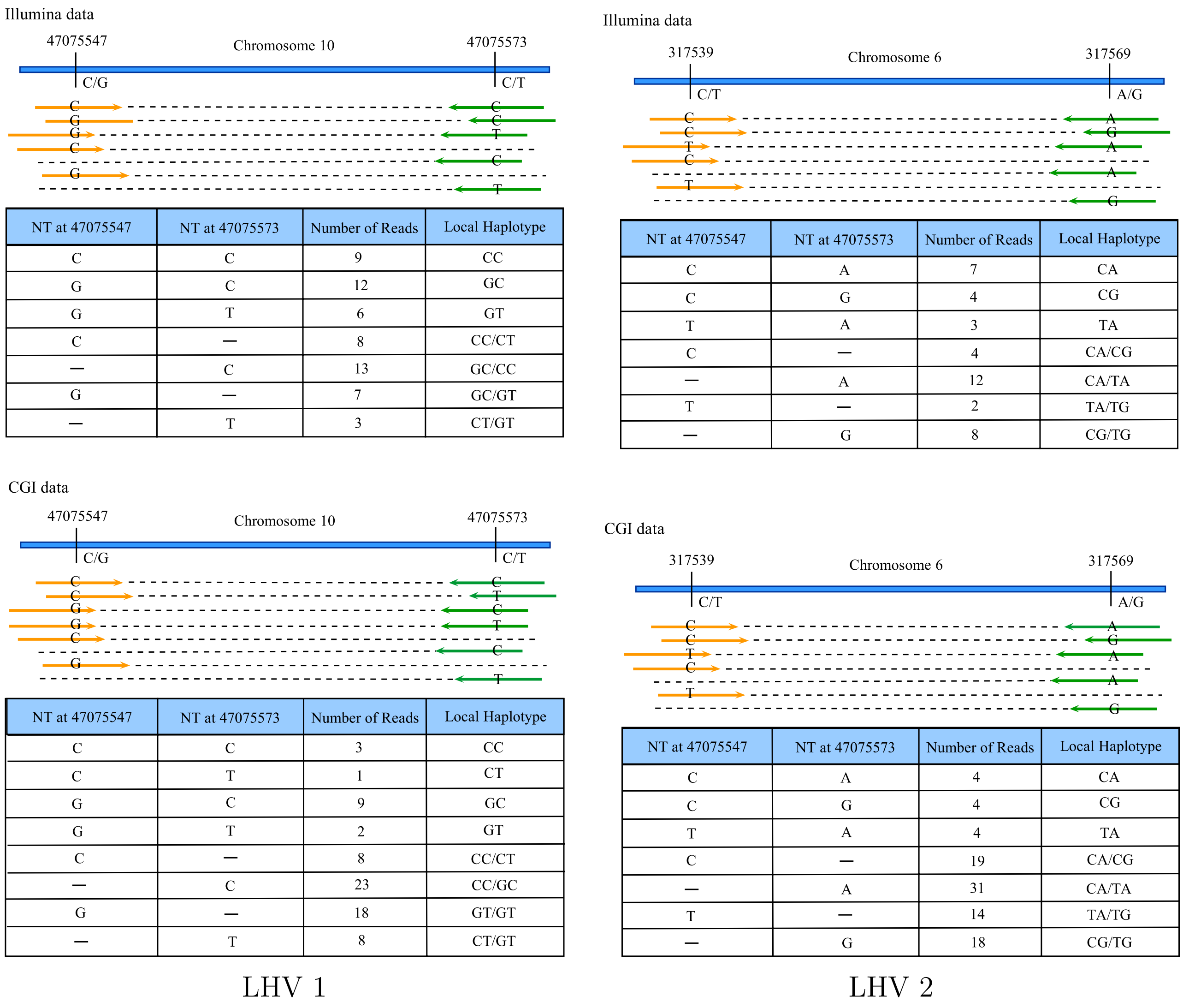


Figure 8: Summary of the LHVs found in the father, mother, and daughter of a family based on both Illumina (ILMN) and CGI data. The ages are 57, 47, and 22, respectively. **Top panel:** A circos plot of prevalence of LHVs for ILMN data. The 3 colored rings describe the genome-wide prevalence and locations of the LHVs for the three family members. Each red triangle dot indicates the existence of at least one LHV in the corresponding genomic region of 1M base pairs. The higher a red dot resides, the larger number of LHVs present in the region. **Bottom panel:** A circos plot of prevalence of LHVs for CGI data. The plot follows the same arrangement as in the Top panel.



Illumina data (Chromosome 10)

NT at 47075547	NT at 47075573	Number of Reads	Local Haplotype
C	C	9	CC
G	C	12	GC
G	T	6	GT
C	—	8	CC/CT
—	C	13	GC/CC
G	—	7	GC/GT
—	T	3	CT/GT

CGI data (Chromosome 10)

NT at 47075547	NT at 47075573	Number of Reads	Local Haplotype
C	C	3	CC
C	T	1	CT
G	C	9	GC
G	T	2	GT
C	—	8	CC/CT
—	C	23	CC/GC
G	—	18	GT/GT
—	T	8	CT/GT

Illumina data (Chromosome 6)

NT at 47075547	NT at 47075573	Number of Reads	Local Haplotype
C	A	7	CA
C	G	4	CG
T	A	3	TA
C	—	4	CA/CG
—	A	12	CA/TA
T	—	2	TA/TG
—	G	8	CG/TG

CGI data (Chromosome 6)

NT at 47075547	NT at 47075573	Number of Reads	Local Haplotype
C	A	4	CA
C	G	4	CG
T	A	4	TA
C	—	19	CA/CG
—	A	31	CA/TA
T	—	14	TA/TG
—	G	18	CG/TG

Figure 9: Two examples of LHVs that overlap between CGI and ILMN data. LHV1 consists of two single nucleotide variants (SNVs) separated by 26 base pairs on Chromosome 10. For the ILMN data (top tables), among all the short reads mapped to this region, 9, 12 and 6 short reads are mapped to both SNVs and exhibit genotypes CC, GC, and GT, respectively. For the CGI data (bottom tables), those numbers are 3, 9, and 2 respectively. Also, many other reads are mapped to one of the SNVs for both data, which reinforces the finding. Statistical inference shows high significance supporting more than two haplotypes in the region. LHV 2 consists of two SNVs separated by 30 base pairs. For the ILMN data, 7, 4 and 3 reads are mapped to both SNVs with genotypes CA, CG and TA, respectively. Those numbers are 4, 4 and 4 for the CGI data.

No. of Variants	Segments with no. of significant haplotypes									Segments with more than 3 variants
	0	1	2	3	4	5	6	7	8	(Not Analyzed)
137886	460	85	2070	90	11	0	0	0	0	457

Table 1: Statistics from *hcf* file of a sample from HNC dataset

No. of Variants	Segments with no. of significant haplotypes									Segments with more than 3 variants
	0	1	2	3	4	5	6	7	8	(Not Analyzed)
6378548	43322	17216	232750	22839	1430	54	2	0	0	196078

Table 2: Statistics from *hcf* file of a sample from CEU trio dataset