Abstract
LARVA — An Integrative Framework for Large-scale Analysis of Recurrent Variants in Noncoding Annotations — And Other Tools for Cancer Genome Analysis
Lucas Sze-wan Fong Lochovsky
2015

Initial approaches to cancer treatment have involved classifying cancer by the site in which it is first formed, and treating it with drugs and other therapies that have very broad targeting. These therapies are often prone to damaging healthy cells in the process, which may lead to additional health complications. With the advent of high-throughput sequencing, and the development of computational tools and software to process the subsequent deluge of sequencing data, much progress has been made on functionally annotating the human genome. Many genomes have been cost-effectively sequenced, providing insight into genetic variation between various human populations. The methods used to study population variation may also be used to study the basis of genetic disease, including cancer. It has now been demonstrated that there are many molecular subtypes of cancer, where each subtype is differentiated based on which important cellular molecule or DNA sequence has been disrupted. Hence, understanding the genetic basis of cancer is paramount to the development of new, personalized molecular therapies to treat cancer.

Noncoding variants are known to be associated with disease, but they are not as commonly investigated as coding variants since assessing the functional impact of a mutation is difficult. For rare mutations, background mutation models have been set up for burden tests to discover highly mutated regions, which might be potential drivers of cancer. This has been developed for coding regions, leading to

the successful use of burden tests to find highly mutated genes. However, this is challenging for noncoding regions because of mutation rate heterogeneity and potential correlations across regions, which give rise to huge overdispersion in the mutation count data. If not corrected, such overdispersions may suggest artefactual mutational hotspots. We address these issues with the development of a new computational framework called LARVA. LARVA intersects whole genome single nucleotide variant (SNV) calls with a comprehensive set of noncoding regulatory elements, and models these elements' mutation counts with a beta-binomial distribution to handle the overdispersion in a principled fashion. Furthermore, in estimating this distribution and determining the local mutation rate, LARVA incorporates regional genomic features like replication timing.

The LARVA framework can be extended in certain ways to facilitate the analysis of its results. By storing information on highly mutated annotations in a relational database, it is possible to quickly extract the most interesting results for further analysis. Furthermore, results from multiple LARVA runs can be combined for a meta-analysis that could involve, for example, finding highly mutated pathways in cancer and other types of genetic disease. Since LARVA's computation consists of many independent units of work, it can benefit from various forms of parallel computation. These forms of computation include distributed computing with a large number of commodity processors, as well as more esoteric types of parallelization, such as general purpose graphics processing unit (GPU) computation.

We make LARVA available as free software tool at larva.gersteinlab.org. We demonstrate the effectiveness of LARVA by showing how it identifies the well-known noncoding drivers, such as TERT promoter, on 760 cancer whole genomes. Furthermore, we show it is able to highlight several novel noncoding regulators that could be potential new noncoding drivers. We also make all of the highly mutated annotations available online.

We also describe the Aggregation and Correlation Toolbox (ACT), a collection of software tools that facilitates the analysis of genomic signal tracks. The aggregation component takes a signal track and a series of genome regions, and creates an aggregate profile of the signal over the given regions. This enables the discovery of consistent signal patterns over related sets of annotations, implying potential connections between the signal and the regions. The correlation component of ACT takes two or more signal tracks and computes all pairwise track correlations. Correlation analyses are useful for finding similarities between various experiments, such as the binding sites of transcription factors as determined by ChIP-seq. The final component of ACT is a saturation tool designed to determine the number of experiments necessary to cover genomic features to saturation. This type of analysis can be illustrated with a ChIP-seq experiment where the inclusion of additional cell lines will reveal more binding sites for a transcription factor of interest: with each new cell line, a smaller fraction of the sites will be newly discovered, and a larger fraction will overlap discovered sites from previously used cell lines. The objective of ACT's saturation tool is to find the point of diminishing

returns in the discovery of new sites, which may result in more efficiently planned experiments.

LARVA — An Integrative Framework for Large-scale Analysis of Recurrent Variants in Noncoding Annotations — And Other Tools for Cancer Genome Analysis

A Dissertation
Presented to the Faculty of the Graduate School
of Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Lucas Sze-wan Fong Lochovsky

Dissertation Director: Prof. Mark B Gerstein

December 2015

Table of Contents

Acknowledgements

I would like to thank:

- My dissertation advisor, Prof. Mark Gerstein, whose continuous guidance and support helped me to carry this research to fruition. Thank you for sharing my excitement for the coming deluge of cancer data—and the resulting newfound possibilities for understanding cancer biology—and for generously maintaining and guiding the Gerstein lab, in which this research was completed.
- The other members of my PhD advisory committee, Prof. Kei Cheung, and Prof. Jing Zhang, for their help and insight.
- The members of the Gerstein lab, especially Dr. Jing Zhang, whose ideas contributed to the final stages of the LARVA project.
- All my friends at CBB, Trinity Baptist Church, and the Yale Graduate Students' Christian Fellowship, for greatly enriching my life outside research, and for creating so many fond memories for me at Yale.
- My brother, Conrad, and my mother and father, Professors Fred and Amelia Lochovsky, whose undying love, support, and numerous e-hugs were instrumental in the completion of this work.

Dedication

To my mother and father,

fountains of unceasing love and encouragement,

and to my God, my heavenly Father,

through whom all things are possible.

1. Introduction

1.1. Motivation

Cancer is the second leading cause of death in the US (Leaf 2004). Despite decades of research and attempts to find the "silver bullet" miracle cure, cancer still kills the same percentage of Americans annually today as it did half a century ago. Thanks to the progress that has been made on heart disease, the current leading cause of death, it is anticipated that cancer will soon take the number one spot, given that its death rate has remained relatively steady year after year. Many ostensible working treatments have been developed, but many have not panned out, ultimately only working on a small fraction of cancers.

The rise of next generation sequencing technologies in the past decade, such as the Illumina and AB SOLiD sequencers (CT 2008), have enabled the whole genome sequencing of a large number of individuals in both a time-effective and cost-effective manner. These technologies assisted in the development of a functional map of the human genome after the Human Genome Project was completed (Birney 2007). Now they are being turned towards the whole genome sequencing of as many cancer patients as possible to understand the molecular basis of cancer disruption processes. We now understand that each cancer, previously classified by the site of the primary tumor, has many different molecular variants, where one particular variant may represent a very small fraction of all patients with that cancer. Combatting cancer will therefore require a precisely targeted molecular therapy for each of these molecular variants.

1.2. Problem Statement

Cancer genomes in general develop a wide range of mutations, owing to the breakage of DNA repair processes that frequently accompanies cancer. Typically, a small fraction of these mutations are actually connected to the breaking of cellular functions that allow cancer to progress. These are known as "driver mutations" (Carter 2009). The rest are considered "passenger mutations", so named because they happen to occur when the cell becomes cancerous, but do not serve any function in making the cell cancerous. Understanding molecular cancer disruptions requires separating the drivers from the passengers.

One of the most common approaches to identifying driver mutations is to identify those mutations that appear frequently and consistently in the sequences of a cancer patient cohort (Parmigiani 2009). Mutations that appear recurrently across many samples more often than would be expected by stochastic mutation processes are likely to be involved in driving cancer progression. This thesis describes the design and development of a software framework built to find recurrent mutations, named LARVA, for Large-scale Analysis of Recurrent Variants in noncoding Annotations. Operating on single nucleotide variant (SNV) data from the whole genome sequence data of cancer patients, LARVA will intersect the SNVs with user-specified annotation sets. These annotation sets can include any region of interest, including genes, pseudogenes, and noncoding RNA, among others. In performing this intersection, LARVA will track the recurrent mutations in the variant dataset, and produce its findings for the user.

Included in the output is a statistical significance test for each recurrently mutated annotation discovered by LARVA. LARVA employs a novel null mutation model that spans the entire genome, and addresses a shortcoming of previous whole genome null models. In previous work (Lohr 2012, Parmigiani 2009, Weinhold 2014), background mutation was modelled as a constant rate over the entire genome. LARVA's model incorporates a variable background mutation rate, which is a more accurate representation of observed mutation patterns. LARVA's model also incorporates the variable mutation accumulation of different genome regions owing to the DNA replication timing of each region in the synthesis S phase of the cell cycle. These two factors are used to produce an expected distribution of variants to determine if the observed recurrent mutations appear at a significantly higher or lower frequency. These significant findings are highlighted for the user, enabling followup analyses to focus specifically on the most likely candidates for cancer progression involvement.

1.3. Thesis Summary and Outline

In this thesis, we describe LARVA—Large-scale Analysis of Recurrent Variants in noncoding Annotations—a computational framework for aggregating rare somatic and germline variants from multiple samples on genomic elements. These *recurrent mutations* serve as a measure of an element's mutation burden, and high-burden elements may correspond to important sites of disruptions for diseases like cancer, and therefore may be crucial for understanding diseases' mechanisms and treatment. LARVA enables the discovery of both recurrent somatic and germline variants in the same annotation, which could implicate previously unknown

disease-causing variants. The following sections explain the concepts of LARVA's framework, and how it functions to identify recurrently mutated genome annotations. This dissertation illustrates how LARVA may be used to study recurrent mutation patterns in both coding regions and noncoding regulatory elements, and sets of pathways and interaction networks. For the purposes of determining if observed recurrent variation is statistically significant, a beta-binomial model of whole genome background mutation is introduced to assess the statistical significance of recurrent variation. This model makes use of a variable genomewide background mutation rate, and the influence of DNA replication timing on the regional background mutation rate, to simulate expected variation across the entire human genome. LARVA's methods have been applied to a set of cancer WGS data, consisting of variants from 760 samples, to demonstrate its usefulness.

This thesis is organized as follows. Section 2 describes previous work relevant to LARVA, recounting earlier uses of whole genome sequencing technology to study population variants, as well as previously developed computational tools with similar design goals, and previous studies involving cancer driver identification and cancer exome studies. Section 3 describes the extensive efforts and procedures that were followed to obtain sets of cancer sequence data suitable for use with LARVA. Section 4 describes early work involving the study of the distribution of cancer single nucleotide variants (SNVs) compared to the variant distribution seen in healthy individuals. Section 5 covers the core concepts behind LARVA, and explains the types of recurrent mutations it is designed to identify. Section 6 provides a detailed description of LARVA and its implementation. Section 7

describes the results that have been obtained on variant data spanning 14 cancers using LARVA. Section 8 describes additional computational tools that may be used to complement a LARVA analysis. Concluding remarks and future directions are given in Section 9.

1.4. Thesis Contributions

LARVA represents a highly optimized method for the discovery of recurrently mutated annotations in any set of SNV calls, spanning any number of samples, across any set of genome annotations. LARVA facilitates the analysis of cancer, and other diseases with a genetic basis, by rapidly identifying the portions of the genome that are consistently broken across many patients. These points of mutation could imply the basis of new molecular therapies that may be used to treat genetic disease. LARVA could also be applied to the analysis of rare germline variants to identify potential connections between these variants and somatic disease variants. If both somatic and germline variants coincide in the same functional elements, the presence of these rare germline variants in individuals could serve as a precursor indication of future disease development.

LARVA also introduces a model of the human genome's rate of SNV acquisition that represents the naturally occurring mutation rate biases present in healthy individuals. This whole genome "null model" of the mutation rate is a novel extension of an exome null model developed for the Broad Institute's MutSig tool (Lawrence 2013). We utilize this model to evaluate if the frequency of a recurrently mutated annotation is a statistically significant enrichment relative to the frequency expected in a healthy genome due to natural variation.

2. Previous Work

2.1. Genome Annotation and Population Variation Studies

2.1.1. Followup to the Human Genome Project

In the wake of the completion of the Human Genome Project (Lander 2001), the primary followup work has concerned the discovery of the portions of the genome that perform some role in the operation of the cell. At the most basic level, this includes the protein-coding exons of the genome, which lead to the production of the cell's workhorses. Beyond these, the genome also contains noncoding elements that regulate the transcription and translation processes that produce proteins. Some of these involve the production of RNA transcripts that bind to other transcripts to control their translation. These include the various classes of noncoding RNA, such as micro-RNA, small interfering RNA (siRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA). Other noncoding elements serve as places where regulatory proteins may bind to promote or inhibit access to a protein-coding region.

Initial genomic element identification studies aimed to look for genome regions under evolutionary constraint. Regions important for proper cellular functioning cannot tolerate mutations to the degree that non-important regions can. Hence, we expect that more constrained regions are likely to correspond to the elements we want to identify. Such regions are said to be under "negative selection", since when these regions are mutated, they are likely not to persist in the population, as said mutations probably break some important function that leaves the carrier at a selective disadvantage (Birney 2007).

2.1.2. Next Generation Sequencing (NGS) Technologies and its Applications

The development of next generation sequencing (NGS) technologies reduced both the time and cost of determining the whole genome sequences of individuals (Chi 2008). Early next generation sequencing technologies, which are now known as second generation sequencing technologies, were first developed by three providers: 454 Life Sciences, Illumina, and AB SOLiD. Although each of these involves unique mechanisms, they are all based upon DNA strand synthesis. Starting with a single strand of DNA whose sequence is to be identified (the query), the biochemical aspect of these methods involves the creation of a complementary DNA strand in a manner that facilitates the identification of the basepairs that were incorporated into the complementary strand. The complementary strand basepairs can then be translated into query strand basepairs.

Voelkerding *et al.* (2009) provide a review of the major NGS companies and their sequencing methods. 454 Life Sciences sequencing technology is based on pyrosequencing, which is a means of synthesizing the complementary DNA strand that involves the release of pyrophosphatase. This reaction will produce a certain intensity of light depending on the identity of the incorporated nucleotide. A photosensor is used to capture these intensities and identify the sequence of basepairs on the complementary strand.

Illumina's sequencing technology, however, involves the use of reversible terminator bases. These bases are unique in that they do not permit the extension of any DNA strand in which they are incorporated. Hence, the strand terminates with these bases, which gives rise to the name. Many copies of the query strand are used

in a complementary strand synthesis procedure that includes both non-terminator and terminator bases. As a result, many partial complementary strands will be synthesized. A dye is used to determine the identity of the terminator base in each experiment. Performed in sufficient quantity, there will be enough partial strands such that each base along the complete strand is represented among the terminator bases. Hence, the terminator base identities, combined with the partial strand lengths, allow the derivation of the complete sequence.

Applied Biosystems' technology has a more unique approach to its sequencing technology. Query sequences undergo complementary strand synthesis using an emulsion polymerase chain reaction (PCR) method, similar to 454 Life Sciences' technology. However, the components of the strand synthesis step are a set of fluorescently labelled di-base probes. Each probe represents each possible two-base permutation, and each emits a distinct fluorescent signal. Hence, the sequence of signals can be used to infer the sequence of basepairs in the query sequence.

A new generation of sequencing technology, driven by single molecule sequencing, have begun appearing on the market around 2010. Two of the first companies involved in developing and manufacturing single molecular sequencing kits are Pacific Biosystems and Helicos Biosciences. Their sequencers operate by filling in the complementary strand of the query sequence, but it is done one basepair at a time. A series of fluorescently labelled nucleotides, as well as polymerase, are used in each reaction round, as well as a labelling molecule that caps the complementary strand and prevents sequence extension. Photosensors

capture the fluorescent signal from the incorporated nucleotide for identification, and then the label is cleaved and washed away, and a new reaction round begins.

A number of applications of next generation sequencing technology accelerated the annotation of the human genome. Chromatin immunoprecipitation (ChIP) is a method for determining the sequences of protein binding sites. Crosslinking proteins when they are bound to DNA sequences leaves the protein and DNA irreversibly bound to each other. Upon digestion of the DNA, the sequences bound by the proteins will be protected, allowing them to be elucidated. Previously, the sequence determination step was accomplished by probing the DNA binding sites against a microarray chip. These chips required bits of sequence from many different parts of the genome to probe against the query sequence derived from the immunoprecipitation step. Replacing the microarray chip with next generation sequencing vastly sped up the immunoprecipitation step. This enabled much higher throughput ChIP experiments, leading to greatly expanded knowledge of various protein-DNA binding sites (Mardis 2007).

Another important application of next generation sequencing is RNA-seq. This method involves the reverse transcription of mRNA transcripts into more stable cDNA molecules, which are then sequenced at high speed with a next generation sequencer. This allowed researchers to capture a snapshot of the genome's transcription, indicating which regions are transcribed, and the quantity of transcription under various conditions (Wang 2009). This drove the discovery of both protein coding regions, and elucidated the presence of regulatory transcripts.

2.1.3. Functional Genome Annotation

The development of these technologies, as well as work performed in projects such as the ENCODE consortium (Birney 2007), spurred the creation of a comprehensive catalog of genome annotations. However, additional work was necessary to connect the genome differences between various persons to their observed individual characteristics. In other words, genome-phenome associations still had to be elucidated.

Connecting genetic variation to phenotype was the goal of the 1000 Genomes (1KG) Project (Durbin 2010). This comprised of work carried out by a consortium of labs that aimed to study and functionally characterize at least 95% of human genetic variants with an allele frequency of at least 1%—the minimum allele frequency necessary for a variant to be considered a polymorphism. The overall study spanned five major population groups from Europe, East Asia, South Asia, West Africa, and the Americas. The 1KG pilot phase consisted of three experimental designs. The first, the trio project, involved high coverage whole genome shotgun sequencing of two families (one from Nigeria, one from Utah) including two parents and one daughter. The second project aimed to sequence a large number of individuals at low coverage (2-6x). The sequenced population spanned 59 unrelated subjects from Nigeria, 60 unrelated subjects from Utah, 30 unrelated Han Chinese from Beijing, and 30 unrelated Japanese from Tokyo. The third project involved exon capture of 8,140 exons from 906 randomly selected genes from 697 subjects spanning 7 populations of African, European, and East Asian origin.

Throughout the 1KG project, high throughput sequencing was applied to the discovery of variants within human populations, spanning single nucleotide variants (SNVs), indels, and structural variants. Researchers also discovered more about the evolutionary conservation (or divergence) of various genome regions. Additionally, the similarities and differences between various human populations throughout the world were elucidated (McVean 2012). The 1KG project also motivated the development of new computational methods to characterize variants, and identify which ones to prioritize for more rigorous characterization experiments. Such tools include the Function-based Prioritization of Sequence Variants (FunSeq) tool (Fu 2014), which is specifically designed to bring together multiple datasets concerning sequence-function relationships, and mark a set of variant calls with the annotations that they overlap, along with other relevant function information. The development of these tools for population variants led to the idea that these same tools could be applied to cancer variants to understand the molecular mechanisms behind cancer, as well as other diseases with a genetic basis.

2.2. Previous Computational Tools

A number of computational tools have been previously developed to facilitate the analysis of the impact of variants on noncoding genome annotations. HaploReg (Ward 2011) is one such tool. Its creators have collected information on noncoding annotations from TRANSFAC (Matys 2003), JASPAR (Mathelier 2013), and protein-binding microarray (PBM) experiments (Berger 2006, Berger 2008, Badis 2009). HaploReg also draws upon various SNP and small indel databases to assist in the functional annotation of input variants, including the 1000 Genomes

Project (Durbin 2010) and dbSNP (Sayers 2010). Variant sets used as input are intersected with HaploReg's annotations using BEDTools (Quinlan 2010), a package of often-used scripts for BED files, or data that can be represented in BED files. HaploReg's output presents the user with information to prioritize variants in the input set according to their functional relevance within their linkage disequilibrium (LD) blocks.

RegulomeDB (Boyle 2012) is a similar system that is geared towards the integration of data from the ENCODE project (Birney 2007) and other sources (Badis 2009, Berger 2006, Berger 2008, Boyle 2010, Bryne 2007, Matys 2003, Pique-Regi 2010, Scharer 2009, Wei 2010) to evaluate variant functionality in regulatory regions. Its datasets include experimentally characterized regulatory regions, ChIP-seq information, chromatin state information, and expression quantitative trait loci (eQTL). RegulomeDB also includes computational predictions of regulatory regions to supplement the experimental evidence. These include the use of DNase-seq (Madrigal 2012) to identify protein-DNA binding sites: such sites represent genome regions with a more exposed chromatin state to allow easier access for protein binding, making these sites more sensitive to DNase I digestion. The authors also conducted their own scan of the human genome for position weight matrices (PWMs) corresponding to known transcription factor (TF) motifs. These motifs were also included in RegulomeDB's data. One final source of regulatory annotations was derived from manual curation of literature sources.

Other previously developed computational tools have focused on facilitating the understanding of cancer disruption by identifying pathways whose

consequences are abrogated in cancer patients. These include two recent systems known as cBio (Cerami 2012) and Multi-Dendrix (Leiserson 2013). cBio starts with variant datasets, and a database of genes and their pathway membership information. The cBio system then identifies those pathways mutated with high coverage and high mutual exclusivity. High coverage refers to the presence of mutations in a large proportion of samples, and high exclusivity means that many of the highly damaging, driver mutations appear in mutually exclusive samples, owing to the sufficiency of mutating just one part of a pathway to nullify its function. Multi-Dendrix extends these ideas by introducing new algorithms to find arbitrary sets of genes that exhibit high coverage and mutual exclusivity of variants, rather than being limited to previously established pathways. GEMINI (Paila 2013) is another general system that manages variant call sets and genome annotation sets through an SQL database, and allows users to formulate their own SQL-based queries over the stored data, allowing a wide range of flexibility for exploring variant data.

2.3. Cancer Driver Identification

Many previous approaches to cancer therapy have revolved around "one size fits all" solutions (Urruticoechea 2010). Aside from the physical location of the tumor, it was assumed that many cancers were fundamentally similar to each other. However, drugs developed to treat these cancers were found to only be effective on certain subsets of patients. Studies of cancer on a genetic level revealed that cancers could be subtyped into versions that each had a different molecular basis. With this new understanding, the focus is now on developing therapies that precisely target

the molecular disruptions specific to each patient in a new form of personalized medicine.

The development of these targeted therapies relies on the identification of mutations that drive cancer progression, known as “drivers”. Only a small fraction of the total set of somatic variants in cancer genomes is drivers. The rest are considered “passengers”. Their occurrence is due to the breakdown of DNA repair processes as the cancer becomes more advanced. As a result, many mutations that would be fixed in normal, healthy cells are acquired. Passenger variants are therefore incidental to cancer disruption processes, not the cause of such processes (Parmigiani 2009).

Separating cancer drivers from passengers has been the focus of much cancer research in the years since the cost-effective sequencing of cancer genomes became possible (Torkamani 2008). One approach involves the use of probabilistic models to simulate the variant distribution assuming everything is mutating at the passenger mutation rate (Parmigiani 2009). These models can be used to determine if a gene’s observed mutation pattern significantly differs from its expected mutation pattern. Driver identification has also been attempted by identifying the biological processes that are disrupted (Vandin 2011, Vandin 2012, Leiserson 2013). This method views a process, driven by multiple genomic elements, as a single functional unit that can be disrupted in multiple places. As a result, individual elements within that process may not be significantly mutated across many samples, but the process as an aggregate of its elements is significantly mutated across samples.

There have also been attempts to classify drivers and passengers based on sequence features that indicate functional significance. Representative of this approach is the Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) tool (Wong 2011). CHASM is designed to classify missense mutations as drivers by looking at the consequences of each mutation to the nucleotide and amino acid context of the surrounding gene/protein. These context features are used in a Random Forest classifier to take advantage of the most predictive features in the final software. CHASM's output includes both an empirical p-value, and a Benjamini-Hochberg-corrected p-value. CHASM was originally developed for deployment on servers, but later received a web-based front end called CRAVAT (http://www.cravat.us).

More recently, new approaches have focused on the idea of using the structure of protein-protein interaction (PPI) networks and other functional networks to identify disease-associated genes. These approaches are based on the expectation that the genes responsible for driving disease phenotype are either close to each other in functional networks, or are connected through some indirect, network-based metric. Chuang *et al.* (2007) devised a method to identify PPI subnetworks that served as predictive markers for metastatic breast cancer. This method starts with single genes that exhibit high differential expression between benign and metastatic breast cancer. The surrounding network is then explored for other genes with marked expression differences between the two types of breast cancer. Chuang *et al*'s method then determines whether the combination of neighboring genes, treated as a subnetwork marker, serves as a better classifier of

benign and metastatic breast cancer. The surrounding network is explored for genes that improve this classification until there is no significant improvement. These subnetwork markers were found to outperform single gene markers in validation tests.

Other methods, such as Köhler *et al*'s (2008), incorporate the global PPI network structure, opening up the potential for any protein in the network to be disease-associated. The Köhler algorithm initiates a random walker at a known disease-associated protein, chosen with uniform probability over all known disease-associated proteins. The walker will, on each iteration, randomly move to a neighboring node (also chosen with uniform probability), or will restart its walk at a known disease-associated protein. Over the long term, the probability distribution of where the walker will be in the network on a given iteration approaches a steady state. This steady state distribution is used to identify those proteins that have many connections to known disease-associated proteins, which are flagged as high priority proteins to study for disease associations.

Vanunu *et al.* (2010) extended the random walk concept by using a disease similarity network and a disease-gene association network along with a PPI network. Using the Online Mendelian Inheritance in Man (OMIM) data on diseases and their associated genes, the authors created a network where nodes are diseases, and edges connect similar diseases. The authors then added disease-protein association nodes and edges, and protein interaction edges. A query disease is investigated by finding those diseases it is similar to, and using the proteins associated with those diseases as prior information for a random network walk over

the PPI portion of the network. Vanunu *et al*'s method was demonstrated to be effective at implicating protein complexes in the causation of various diseases, including prostate cancer, Alzheimer's, and diabetes.

2.4. Cancer Exome Studies

So far, many studies of the genetic causes of cancer have been focused on studying the exome features of cancer cells. The focus of this section is to highlight the findings of a few of these studies.

Grasso *et al.* (2012) sequenced the exomes of some 50 lethal, heavily pre-treated metastatic castration-resistant prostate cancers (CRPCs), and 11 treatment-naïve, high-grade localized prostate cancers. The sequence data generated was then used to study a range of mutation classes, including SNVs, insertions, deletions, and copy number variants (CNVs). The analyses drew attention to the roles of CHD1 and ETS deletions in prostate cancer. Also important was the discovery of recurrent mutations in chromatin- and histone-modifying genes, which interact with androgen receptor (AR), which was previously demonstrated to drive prostate cancer progression (Shen 2010). Other genes known to interact with AR were also found to be mutated, including FOXA1, MLL2, UTX, and ASXL1.

A larger study was conducted by Barbieri *et al.* (2012) that spanned 112 prostate tumor/normal sample pairs. Exome sequences were used to identify recurrent variants and recurrently mutated genes. Most notable in the findings was the discovery of recurrent SPOP, MED12, and FOXA1 mutations (thus backing up the Grasso analyses). The authors put extra emphasis on exploring the SPOP mutations, as these had been previously reported in prostate cancer, but not fully understood

on a functional level. It was discovered that SPOP's mutations affect conserved residues in the substrate binding cleft.

Krauthammer *et al.* (2012) studied the exomes of some 147 tumor/normal melanomas. They split these melanomas by the amount of sunlight shining on the tumor site (i.e. sun-exposed vs. sun-shielded), and determined the mutation landscape of each melanoma type. The sun-shielded samples were further segregated by primary tumor site, which included acral, mucosal, and uveal melanomas. Sun-exposed melanomas had far more ultraviolet somatic mutations, as expected. The two most frequently mutated genes were *BRAF* and *NRAS*, confirming previous melanoma driver research (Ribas 2011, Jakob 2012). Newly discovered from this analysis was an activating mutation in *RAC1*, appearing with the third-highest frequency of all recurrent mutations in sun-exposed melanomas. The authors also discovered a number of other genes mutated at a lower frequency in both sun-exposed and sun-shielded melanomas.

3. Assemblage of Cancer Data

3.1. Special Challenges with Obtaining Data from Protected Data Sources

Cancer data that can be linked to the contributing individuals, either directly or by joining multiple datasets on a foreign key relationship, is considered protected data. Prospective users of this data are required to submit applications detailing the specific uses of the data in their research, and indicate that their IT systems meet the requirements of storing the data securely. We have navigated the protected data access applications and incidental procedures of several cancer data repositories, including The Cancer Genome Atlas (TCGA)(Muzny 2012, Bell 2011), the Database of Genotypes and Phenotypes (dbGaP)(Sayers 2010), the International Cancer Genome Consortium (ICGC)(Hudson 2010), and the European Genome-Phenome Archive (EGA)(Leinonen 2010). Following is a description of the requirements we encountered on the path to obtaining protected data access.

Protected data access applications require a description of the research in which the requested data will be used. In some cases, a layman's description is also required for the purposes of demonstrating the usefulness of the data to the general public and to benefactors. A list of relevant publications is required to demonstrate the experience and pedigree of the applying lab. Individuals who will be granted access to the data must also be listed. Access renewal is typically required every year, with updates on research progress and future plans.

Additionally, applicants' computational infrastructure must support certain safeguards to prevent unauthorized access to the data while it remains in the possession of the applicants. Computer systems containing the data, and any backup

copies, must be physically secure, typically by having them under lock and key. Logins must be controlled so that only the individuals named on the access application can access the data, and any network access to the data must be secure from external intrusion. Any copies of the data taken onto portable devices, including laptops, smartphones, and tablets, must be encrypted due to the greater risk of data falling into the wrong hands. Furthermore, when all research using the protected data is complete, all copies must be destroyed. Exceptions are allowed when the data must be archived to comply with national audits or legal requirements. Finally, everyone working with the data must be trained in the responsible use of confidential patient data, and the security protocols established around the data.

The use of data involving human subjects is, in most cases, subject to the approval of an institutional review board (IRB), which must examine the proposed research plan and verify that the human subjects and their data are treated with proper ethical standards. However, human subject research only requires IRB approval if the research involves collecting data on human subjects through intervention or interaction with the individual, or data that can be used to personally identify the individual it was collected from. Since these conditions did not apply to the research described in this dissertation, IRB approval was not required.

3.2. Manifest of Cancer Data Collected

The first data portal for which we applied for protected access was The Cancer Genome Atlas (TCGA)(Bell 2011, Muzny 2012). TCGA is a collaborative effort

between multiple labs and groups located throughout the US, Canada, Europe, and Australia. Our initial examination of the Data Portal revealed that there were a number of whole exome sequence variant datasets available, which are described in Table 1.

Table 1: List of exome cancer datasets obtained from TCGA.

Cancer	Protocol	Center	Sequencer	# Tumor/ Normal Samples	Reference
Colon adenocarcinoma	WXS	BCM	Illumina	52	Muzny 2012
Colon adenocarcinoma	WXS	BCM	SOLiD	53	Muzny 2012
Ovarian serous cystadenocarcinoma	WXS	BCM	SOLiD	91	Bell 2011
Ovarian serous cystadenocarcinoma	WXS	WUSTL	Illumina	88	Bell 2011
Rectum adenocarcinoma	WXS	BCM	Illumina	12	Muzny 2012
Rectum adenocarcinoma	WXS	BCM	SOLiD	35	Muzny 2012

Later, we were granted access to the TCGA protected data stored in the Database of Genotypes and Phenotypes (dbGaP)(Sayers 2010). This data can be searched and downloaded using CGHub (https://cghub.ucsc.edu/index.html), a utility built by the TCGA to facilitate bulk downloading of TCGA data. Using CGHub, we obtained additional data, including TCGA's prostate cancer RNA-seq and whole genome sequence (WGS) data (Table 2).

Table 2: List of protected cancer datasets obtained from TCGA via CGHub.

Cancer	Protocol	# Tumor/Normal Samples
Prostate cancer	RNA-seq	497
Prostate cancer	WGS	20
Ovarian serous cystadenocarcinoma	WGS	8
Kidney carcinoma	WGS	32

In addition to this, we have initiated a number of collaborations with groups that have directly sequenced cancer patients. This has given us access to the cancer data listed in Table 3.

Table 3: List of cancer datasets obtained from collaborators

Cancer	Source Lab	Protocol	# Samples	Reference
Prostate cancer	Dr. Mark Rubin	WGS	7	Berger 2011
Glioma	Dr. Murat Günel	WGS	26	[unpublished]
Melanoma	Dr. Ruth Halaban	WXS	316	Krauthammer 2012
Medulloblastoma	Dr. Jan Korbel	WGS	3	Rausch 2012
Prostate cancer	Dr. Jan Korbel	WGS	11	Weischenfeldt 2013

The remainder of the cancer data we have worked with was derived from a review of major cancer publications at the time. These datasets are listed in Table 4.

Table 4: List of cancer datasets obtained from publications

Cancer	Source Lab	Protocol	# Samples	Reference
Breast cancer	Dr. Michael Stratton	WGS	21	Nik-Zainal 2012
Prostate cancer	Dr. Levi Garraway	WGS	57	Baca 2013
Prostate cancer	Dr. Scott A. Tomlins	WXS	61	Grasso 2012
Prostate cancer	Dr. Levi Garraway	WXS	112	Barbieri 2012
Malignant melanoma	Dr. Michael Stratton	WGS	1	Pleasance 2009
Lung cancer	Dr. Matthew Meyerson	WXS	183	Imielinski 2012
Stomach cancer	Dr. Mao Mao and Dr. Suet Yi Leung	WGS	100	Wang 2014
Breast cancer	Dr. Michael Stratton	WGS	119	Alexandrov 2013
Lung adenocarcinoma	Dr. Michael Stratton	WGS	24	Alexandrov 2013
Acute Myeloid Leukemia	Dr. Michael Stratton	WGS	7	Alexandrov 2013
Acute Lymphoblastic Leukemia	Dr. Michael Stratton	WGS	1	Alexandrov 2013
Chronic Lymphocytic Leukemia	Dr. Michael Stratton	WGS	28	Alexandrov 2013
Pancreatic cancer	Dr. Michael Stratton	WGS	15	Alexandrov 2013
Pilocytic Astrocytoma	Dr. Michael Stratton	WGS	101	Alexandrov 2013
Medulloblastoma	Dr. Michael Stratton	WGS	100	Alexandrov 2013
Liver cancer	Dr. Michael Stratton	WGS	88	Alexandrov 2013
Lymphoma B-cell	Dr. Michael Stratton	WGS	24	Alexandrov 2013

4. Comparison of Population Variation Patterns with Cancer Variation Patterns

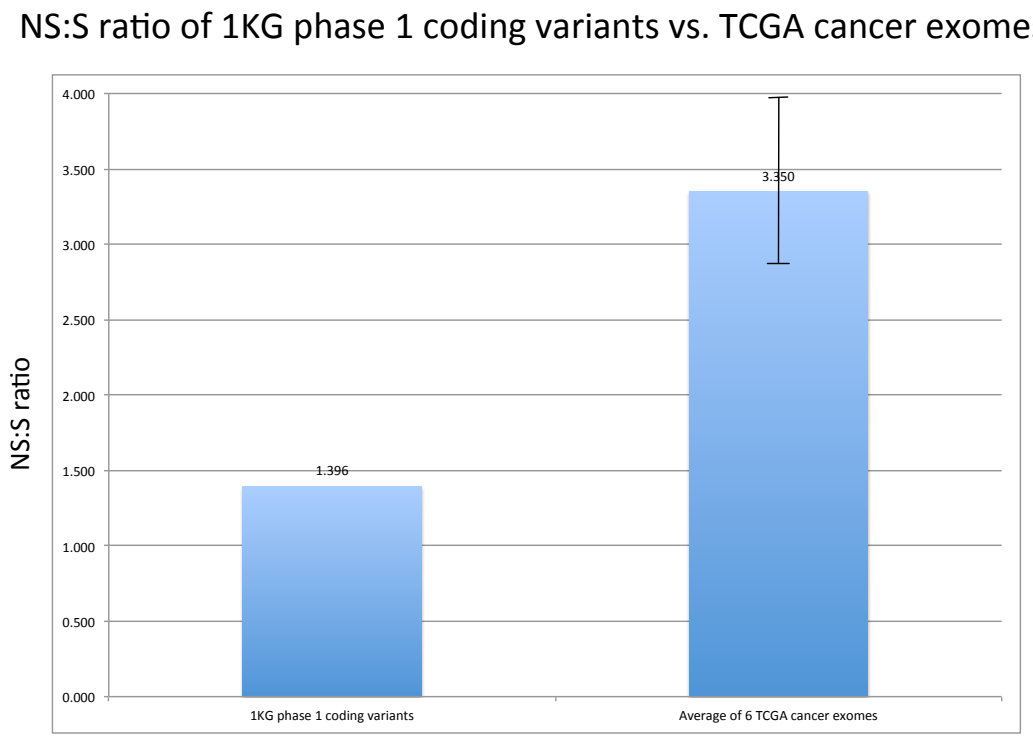
At the time of this analysis, the cancer exome data available spanned the TCGA exomes from Table 1, and the melanoma exomes provided by Dr. Ruth Halaban (Table 3). With this data, the nonsynonymous:synonymous (NS:S) ratio of each cancer was compared to the NS:S ratio of the 1000 Genomes Project (1KG) phase 1 samples. An intersection analysis was also conducted between the cancer exome variants and the 1KG phase 1 variants. Each cancer's variants were intersected with the 1KG coding variant set. The cancer variants were also divided into drivers and passengers using the Cancer-Specific High-throughput Annotation of Somatic Mutations (CHASM)(Wong 2011), a computational classification engine that separates nonsynonymous variants into driver and passenger groups using multiple relevant features. The 1KG coding set was also divided into common and rare variants (according to the derived allele frequency (DAF)), and into nonsynonymous and synonymous variants. These sets, along with the set of all 1KG variants, resulted in a total of five 1KG variant sets. Each of these sets was intersected with the cancer exome driver and passenger sets to identify enrichments and depletions of drivers and passengers in each 1KG variant set.

Enrichments and depletions were determined by comparing the observed intersection with an expected intersection computed by simulation. These simulations involved randomizing the positions of the variants in each cancer dataset 10,000 times, producing 10,000 sets of random variants for each cancer. Each of these random datasets was intersected with each of the five 1KG coding

variant sets. Hence, a distribution of intersections expected at random was created. These distributions were used to determine any significant enrichments or depletions of cancer variant intersections with the 1KG coding variants.

Comparison of the cancer nonsynonymous:synonymous (NS:S) ratio to that of the 1KG phase 1 variants indicates that there is an almost threefold difference (Fig. 1). This finding indicates that cancer variants hit sites with a significant protein coding effect much more than sites that are silent in the protein code. Given cancer's disruption to important growth and regulatory processes, this is a logical expectation.

Fig. 1: The ratio of nonsynonymous to synonymous variants in the phase 1 protein coding variants of the 1000 Genomes Project, compared to the ratio seen in variants derived from the exomes of two studies of colon adenocarcinoma, two studies of ovarian serous cystadenocarcinoma, and two studies of rectum adenocarcinoma.



The intersection analysis with the Halaban melanoma variants (Krauthammer 2012) and 1KG coding variants (Table 5) indicates that there is no significant enrichment among cancer drivers. This means that there is a significant non-overlap between cancer drivers and 1KG coding variants, which is expected, since the healthy samples in 1KG should have cancer if that were not true. Among the cancer passenger variants, there is a significant enrichment, which makes sense given that cancer often disrupts DNA repair mechanisms, allowing everything to mutate randomly. When the 1KG variants are split into common and rare variants, it is clear that cancer passengers are enriched in both groups, but the enrichment ratio is much higher in rare variants. This is probably reflective of the fact that rare variants likely represent risk alleles that serve as precursors to genetic disease, hence mutation of rare variants is favored for driving cancer processes. The intersection analysis between cancer variants and nonsynonymous/synonymous 1KG variants indicated that there was an enrichment of cancer variants in both 1KG categories, but there was no significant difference in the enrichment between the two groups.

Table 5: Melanoma variants obtained from the lab of Dr. Ruth Halaban were intersected with variants derived from the 1000 Genomes Project sequence data to identify significant enrichments or depletions of somatic cancer variants in germline variants. P-values and confidence intervals were determined by comparison to an expected distribution of intersecting variants derived by random variant simulation. Drivers are neither enriched nor depleted in 1KG coding variants. Passengers are enriched in 1KG coding variants, with a clear difference in the enrichment ratios of common and rare variants.

Drivers intersection with	**Observed count**	**Average random count**	**Observed/Random Ratio**	**p-value**
1KG coding variants	0	0.0150	0	0.452
1KG coding common variants	0	0.0027	0	0.479
1KG coding rare variants	0	0.0106	0	0.459
1KG coding nonsyn variants	0	0.0093	0	0.462
1KG coding syn variants	0	0.0057	0	0.470

Passengers intersection with	Observed count	Average random count	Observed/Random Ratio	p-value
1KG coding variants	3924	543.199	7.22	0
1KG coding common variants	638	110.720	5.76	0
1KG coding rare variants	2866	388.141	7.38	0
1KG coding nonsyn variants	2245	313.112	7.17	0
1KG coding syn variants	1708	230.104	7.42	0

At the time of this WGS analysis, there were three WGS cancer datasets available from Berger *et al.* (2011), Pleasance *et al.* (2009), and Nik-Zainal *et al.* (2012). With this data, the sample-normalized fraction of variants in each dataset that fall within GENCODE v7 (Harrow 2012) genes, pseudogenes, noncoding RNA (ncRNA) regions like microRNA (miRNA) sites, and transcription factor binding sites (TFBSes) was calculated. These fractions were compared to the corresponding fractions for NA12878, one of the deeply sequenced trios from the 1KG project's pilot phase (Durbin 2010). Table 6, which shows the percent of variants from each whole genome sequenced (WGS) cancer dataset that map to major genomic regions, indicates differences in cancer variant distribution compared to the normal distribution of NA12878. A larger proportion of cancer variants map to genes and pseudogenes, while a smaller proportion map to TFBSes. The regions with larger proportions of cancer variants indicate that cancer targets coding regions for disruption relatively more often than regulatory targets.

Table 6: The percent of variants from various WGS cancer datasets in four major classes of genome annotation, with NA12878 serving as a normal reference. Enrichment and depletion observed in these categories indicates the targets of cancer disruption.

Variant data	Genes (Exons)	Pseudogenes	RNA	TFBSes
NA12878	0.664%	0.314%	0.125%	16.4%
Prostate Cancer (Berger 2007)	0.653%	0.386%	0.115%	12.2%
Melanoma (Pleasance 2009)	0.879%	0.360%	0.084%	11.9%
Breast Cancer (Nik-Zainal 2012)	0.944%	0.390%	0.121%	12.9%

5. Introduction to LARVA

5.1. Preamble

Genomes of numerous patients have been sequenced (Almasy 2014, Baca 2013, Barbieri 2012, Grasso 2012, Shi 2013) opening up opportunities to identify the underlying genetic causes for complex disease (Chen 2014, Stefansson 2014, Tervasmäki 2014, Zhang 2014) and develop more effective therapies targeted at specific molecular disease subtypes (Kurtova 2014). Most of these studies have so far focused on identifying mutations and defects in the protein coding regions, or exomes, of disease genomes (Baca 2013, Lawrence 2013, Long 2014, Rudd 2014, Yadav 2014). These methods usually search for coding regions with higher than expected mutation frequencies in protein coding genes through rigorous background mutation rate control over a variety of genomic features (Lawrence 2013). Such methods have been successfully used on numerous cancer genomes (Youn & Simon 2011). However, the noncoding regions, which comprise more than 98% of the human genome, were rarely investigated, primarily due to the difficulty of functional interpretation of noncoding variants.

Recent genome annotation analysis has revealed that a significant portion of the human genome is functional in a certain tissue or development stage (Dunham 2012, Gerstein 2014), and several noncoding variants have been implicated in disease (Fu 2014). For example, several genome-wide association studies (GWAS) studies have discovered the phenotypic effect of common noncoding variants in regulatory regions (Dees 2012, Futreal 2004). Other studies have reported that

noncoding TERT mutations drive cancer progression in multiple tumor types, including melanomas and gliomas (Grossman 2013, Maurano 2012, Vinagre 2013). Moreover, mutations in the promoter regions of PLEKHS1, WDR74 and SDHD were also identified as recurrent driver mutations in some cancer types (Weinhold 2014). In another example, analysis of the miRNA-binding sites on BRCA1 and BRCA2, the established risk genes of breast cancer, indicated that certain variants in these sites are associated with increased likelihood of early onset breast cancer (Erturk 2014). Furthermore, some references showed that a histone H1 variant is linked to oncogene expression in ovarian cancer (Medrzycki 2014). In light of these discoveries, and the growing availability of whole-genome sequencing data (Alexandrov 2013, Baca 2013, Berger 2011, Cancer Genome Atlas Research 2013, McLendon 2008, Wang 2014, Weischenfeldt 2013), a statistical framework facilitating the identification of highly mutated noncoding mutations is called for.

More recently, a genome wide computational effort has been made to discover the noncoding regions with higher mutation burden in cancer genomes (Weinhold 2014). The authors called whole genome somatic variants for 863 human tumor sequences from The Cancer Genome Atlas (TCGA) (McLendon 2008), and analyzed the variants that fall into noncoding annotations. A p-value was computed for each annotation reflecting the likelihood that the given annotation had more variants than expected from background mutation processes, which was modelled with a binomial distribution. They successfully identified some known noncoding drivers, such as the TERT promoter, and reported some novel candidates that were not discovered previously. The use of the binomial distribution is based on two

assumptions: 1) the mutation rate is homogeneous; 2) variants arise independently. However, cancer genomes often violate these assumptions. First, studies on the coding variants already proved that the mutation rates in cancer genomes demonstrate substantial cancer type, sample, and regional heterogeneity (Lawrence 2013). Second, some passenger mutations were generated by other driver events, such as structural alterations and mutations in DNA replication or repair genes (Hodgkinson 2011). In the human genome, there are many regions with highly correlated mutational profiles. For instance, the germline variant distribution is influenced by the high linkage disequilibrium (LD) of many regions, and for somatic variants, there are many known hotspots. Hence, some degree of dependency is to be expected in the human germline and somatic mutation landscape. Consistent with these statements, we observed that the somatic mutation counts in the noncoding elements exhibited substantially higher variance than expected, or overdispersion, indicating that a binomial distribution might be potentially inadequate to handle such data, and the resultant p-values might be heavily inflated. Hence, if this p-value inflation is not taken care of, a significance calculation based on a binomial distribution might report some artificial mutation hotspots by chance instead of real driver events.

Sections 5-7 present a computational system, LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations), that identifies highly mutated noncoding regulatory elements using whole genome sequencing (WGS) variant data from multiple genetic disease patients. LARVA treats the mutation counts within a given regulatory element as a beta-binomial distributed random variable. This

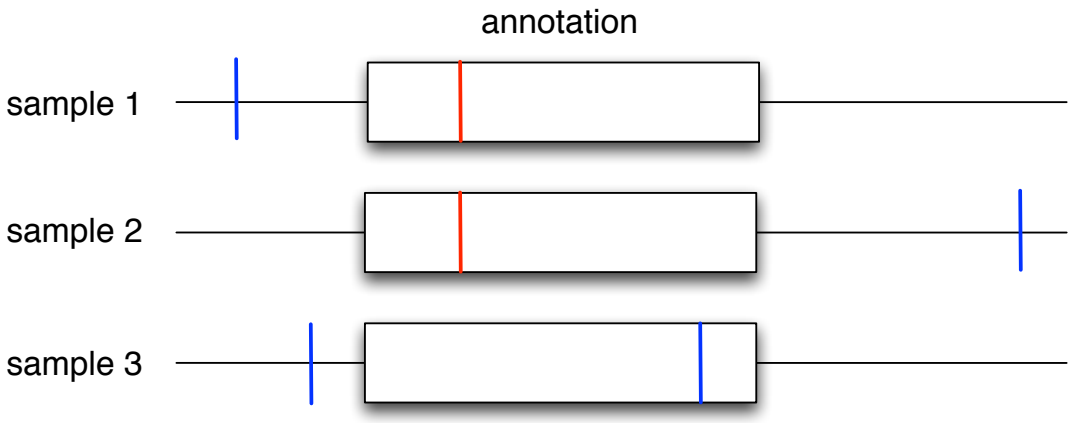
design automatically accommodates the heterogeneous nature of mutation accumulation in cancer genomes and the potential dependency among neighboring loci by allowing the local mutation rate to be drawn from a beta distribution. Furthermore, we also divided the whole genome into several local bins and classified them using some known genomic confounders of the mutation rate, such as replication timing, for a more accurate local background mutation model. Such integrative analysis could potentially control the false positive rate in an effective manner. We demonstrate the usefulness of LARVA for finding both well-known and novel noncoding regulators with higher mutation burdens in a set of WGS cancer data that represents all the different types of whole genome sequenced cancers to our knowledge (see section 7 for details). We release the noncoding annotations, the mutation counts, and the corresponding p-values on the 760 cancer genomes used in this dissertation as a potentially powerful resource to facilitate cancer researchers for driver events discovery and validation in the future. Although designed for somatic variant analysis, the logic of LARVA can be immediately extended for germline variant analysis in complex diseases. The following sections describe LARVA's concepts, their applications to the study of genetic disease, and cancer findings derived with LARVA. LARVA was published in *Nucleic Acids Research* (Lochovsky 2015).

5.2. Recurrent Variants

Recurrent variants here mean SNVs from multiple samples that overlap (i.e. they have the same coordinate). In Fig. 2, which is a simple example illustrating the SNVs in three samples around the same annotation, the annotation contains variants

from all 3 samples, but the variants from samples 1 and 2, highlighted in red, overlap perfectly. Such mutations may correspond to a critical component of a gene's product that is important for tumor suppression. These mutations may also be used to classify the subtype and severity of cancer patients (Vandin 2011).

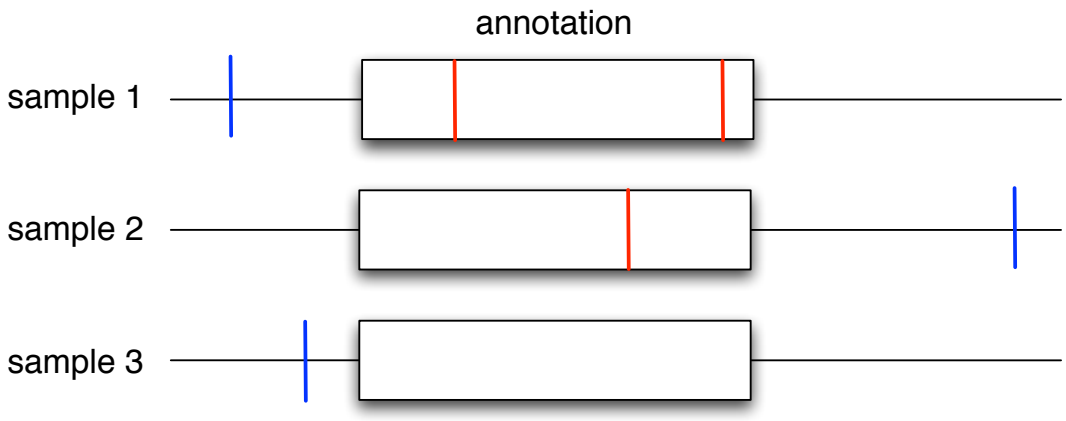
Fig 2: Recurrent variants are single nucleotide variants (SNVs) from multiple samples that overlap in a single annotation.



5.3. Recurrently Mutated Annotations

Recurrently mutated annotations refer to annotations that contain SNVs from multiple samples that do not necessarily overlap. Such annotations may be functionally disruptable in multiple places, and therefore, multiple patients with the same functional disruption may carry SNVs in different places of the same gene. Hence, it is important to detect whether the annotation itself is mutated in multiple cancer patients, rather than individual positions. Fig. 3 illustrates an annotation recurrently mutated in two samples, with the relevant variants highlighted in red. Unlike Fig. 2, these variants do not overlap.

Fig 3: Recurrently mutated annotations contain variants from multiple samples that are positioned anywhere within the annotation boundaries.



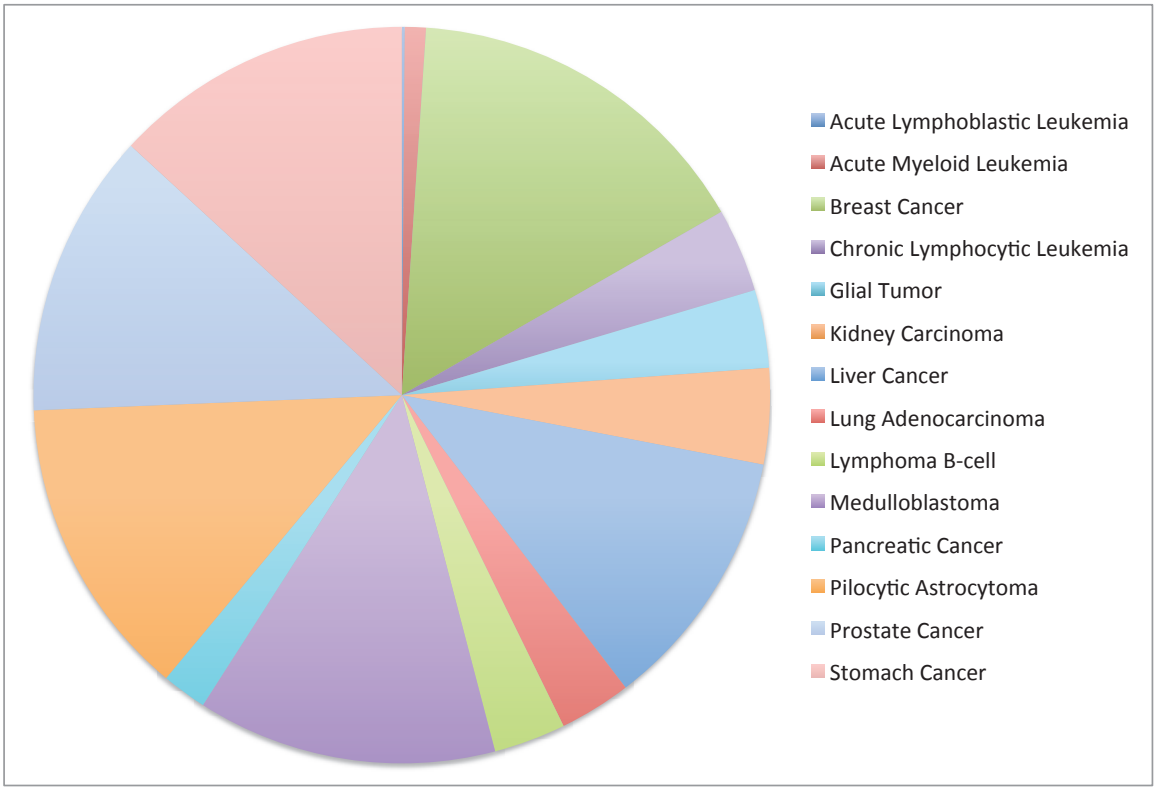
The focus of LARVA is to identify noncoding regulatory annotations that are recurrently mutated. The following section describes the collection of WGS cancer variants, noncoding annotations, and LARVA's methods for finding significant mutation burdens in those annotations.

6. LARVA Data and Implementation

6.1. Whole genome cancer variant data

We collected whole genome cancer variant calls from a large number of previously sequenced cancer genomes. The majority of our data came from a set of 507 whole genome cancer samples published in Alexandrov et al. (2013). This data spans breast cancer, lung cancer, leukemia, pancreatic cancer, pilocytic astrocytoma, medulloblastoma, liver cancer, and lymphoma (Fig 4 and Table S1). This was supplemented with a collection of 95 prostate cancer samples we obtained from publications (Baca 2013, McLendon 2008, Berger 2011, Weischenfeldt 2013), a set of 26 unpublished glial tumor samples, 32 kidney cancer samples from the TCGA (McLendon 2008), a set of 100 stomach cancer samples from Wang *et al.* (2014).

Fig. 4: Distribution of WGS cancer samples obtained for LARVA analysis.



6.2. Quality control of the WGS variants

A number of genomic regions are known to have poor read mappability due to sequence phenomena that cause ambiguous mapping results, such as a large number of tandem repeats. These regions are known as signal artifact blacklist regions (Derrien 2012). Since it is likely that variant calls in this region are possibly inaccurate, we opted not to use these regions or any intersecting variants in our mutation rate calculations (details in Fig. S1). Blacklist regions were derived from Derrien *et al.* (2012), and downloaded from the UCSC Genome Browser. Variants intersecting these regions, as determined by BEDTools (Quinlan 2010), were removed from the analysis.

6.3. Noncoding annotation summary

Our analysis covered a range of noncoding regulatory annotations. The GENCODE v16 main annotation file was parsed to derive the coordinates of regulatory annotations close to gene regions, including promoters and untranslated regions (UTRs)(Harrow 2012). Transcription factor (TF) binding sites were derived from the Chip-seq experiments conducted as part of the ENCODE project (Rozowsky 2009). We collected the full list of TF binding sites in all possible tissues and cell lines from ENCODE. Distal regulatory modules (DRM) enhancers, which regulate the expression of genes at non-adjacent sites, were derived from (Yip 2012). Another class of regulators, the Dnase I hypersensitive (DHS) sites (Thurman 2012), were also derived from the ENCODE project. Additionally, we added a set of sites deemed "ultra-conserved" in Bejerano (2004) due to their extremely high level of conservation across many species. Furthermore, we used a set of "ultra-sensitive"

sites from Khurana *et al.* (2013), so named because they are noncoding regions under higher selective pressure from the population genetics perspective. Finally, similar to the 2500bp promoter sites, we studied the more proximal transcription start sites (TSSes) by extracting the 100bp regions immediately upstream of GENCODE gene coding annotations (Harrow 2012). Table 7 summarizes the noncoding annotations.

Table 7: A summary of the noncoding annotations of interest

Feature	Nucleotide	Number	Mean Length	Length SD	Feature Genome Fraction
Promoters	89325819	72965	2500	0	0.028473455
TF Peaks	376580899	5710734	632.3407	434.9489	0.120038744
DHS Sites	434080020	2890742	150.162145	4.519059	0.138367136
DRM Enhancers	8273100	9599	861.871	987.8922	0.002637129
Ultra-conserved sites	126007	481	261.96881	70.47657	4.02E-05
Ultra-sensitive sites	610048	1354	683.1492	769.1901	0.000194459
UTRs	41398790	155052	392.5388	781.4613	0.013196258

Pseudogenes are known hotspots for artifacts due to their high context resemblance to their parent genes. In order to avoid potential variant calling bias, partially due to mapping difficulty, we removed the promoters, TSS, and UTR analyses for pseudogenes in the GENCODE annotation (details in Fig S2).

6.4. Models used for significance evaluation of mutation burden

The mutation counts for each regulatory element were calculated from the 760 cancer genomes mentioned above. For each regulatory element category, three models were used to calculate the mutation rate that would be expected due to background stochastic mutation processes for significance evaluation.

Suppose there are k noncoding regulatory elements (e.g. TF binding sites) to be analyzed. For the i^{th} element, let n_i stand for the total number of nucleotides in i. x_i and p represent the number of mutations within element i and the probability of observing a mutation in each position. Some previous models (Ding 2010, Weinhold 2014) assumed that p is constant over the entire genome and mutations occur in an independent way. Hence, in model 1 x_i can be described as a binomial distribution:

$$x_i : Binomial(n_i, p) \tag{1}$$

However, due to the heterogeneous nature of the cancer genomes and the possible dependencies among neighboring loci, large overdispersion was found in the mutation count data (as seen in Fig. 8 in section 7.3). As a result we first improved model 1 into a two-layer hierarchical model (model 2). Instead of setting p as a constant, we allow it to be drawn from a beta distribution with two parameters μ and σ indicating the average mutation rate and overdispersion respectively (details in Supplemental section 2.2). As a result, the marginal distribution of x_i follows a beta-binomial distribution:

$$\begin{aligned} x_i \,|\, p_i &: Binomial(n_i, p_i) \\ p_i &: Beta(\mu, \sigma) \end{aligned} \tag{2}$$

Furthermore, mutation rates are known to be confounded by a lot of genomic features, such as replication timing (represented by R), so we further divided the noncoding regulatory elements into 10 bins according to the averaged replication timing signal. Within each bin, we assumed that the mutation rate follows the same distribution. Therefore, model 3 can be represented as:

$$\begin{aligned} & x_i \mid p_i : Binomial\left(n_i, p_i\right) \\ & p_i : Beta\left(\mu \mid R, \sigma \mid R\right) \\ & \mu \mid R, \sigma \mid R : \text{constant within the same } R \text{ bin} \end{aligned} \tag{3}$$

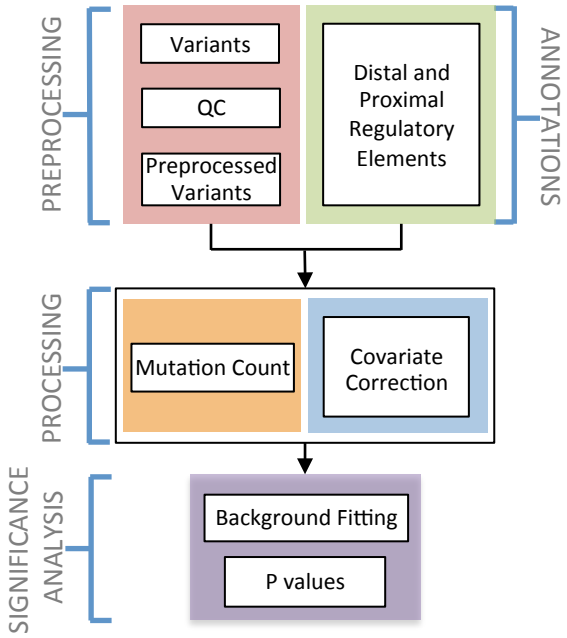
Maximum likelihood estimation was used for model 1. The moment estimator mentioned in (Kleinman 1975, Young-Xu 2008) was used to estimate the parameters in models 2 and 3, and the p-values were calculated accordingly for the three models (for details see Supplemental section 2.2).

6.5. Workflow of LARVA

The workflow of LARVA is given in Fig. 5. The cancer variants in VCF format pass through a quality control filter that includes removing those variants that fall into blacklist regions. The preprocessed variants, along with our collected set of noncoding annotations that do not overlap blacklist regions, are used in the main computation. The main processing step includes counting all variant intersections with the noncoding annotations. DNA replication timing was used in model 3 for local mutation rate corrections. For each annotation category, the background

mutation model was calculated using models 1-3 mentioned above, and p-values were given accordingly.

Fig 5: A flowchart of LARVA's procedure for identifying significant highly mutated noncoding elements. Cancer variants in VCF format are passed through quality control filters, and then intersected with our noncoding annotation corpus. After factoring in regional mutation rate corrections, a beta-binomial distribution is fitted to the observed data, which allows the identification of elements with a significant mutational burden.



6.6. Release of results

We release the noncoding annotations, the mutation counts, and the corresponding p-values on the 760 cancer genomes used in this paper as a potentially useful resource to facilitate cancer researchers for driver event discovery and validation in the future. The files can be directly downloaded from larva.gersteinlab.org. The files available for download include:

- C++ source code with documentation and a regression test suite

- A LARVA Docker image, which encapsulates all of LARVA's prerequisite software and greatly simplifies installation
- Our noncoding annotation collection, and
- Our p-values from running LARVA with our cancer variant collection on our noncoding annotation collection

7. LARVA Cancer Results

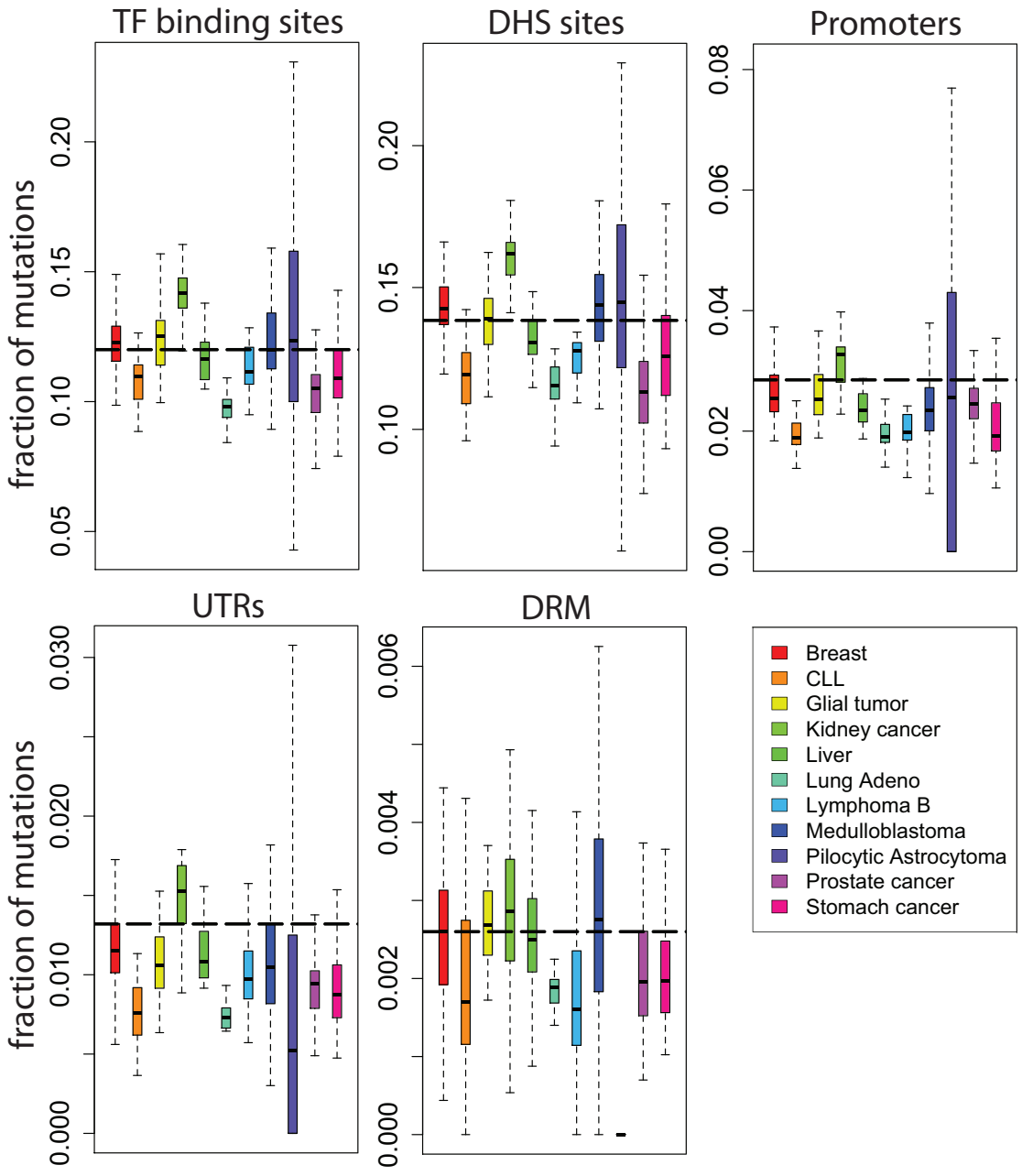
7.1. Overview of the annotated noncoding variants on various cancer genomes

We sought to study the whole genome somatic mutation patterns of as many different cancer patients as possible. To that end, we collected whole genome cancer variant call sets from a range of cancer data repositories (Alexandrov 2013, McLendon 2008) and publications (Alexandrov 2013, Baca 2013, Berger 2011, Cancer Genome Atlas Research 2013, Wang 2014, Weischenfeldt 2013). Our data spans 760 genomes, and includes 14 types of cancer (Fig 4 and Table S1).

As shown in Table 7, our noncoding annotation list spans approximately 30% of the human genome. We observed different cancer types demonstrate distinct mutational preferences over these noncoding regions. To illustrate this phenomenon, we used 11 types of cancer from our overall dataset for which there are at least 20 samples and calculated the fraction of WGS mutations within each noncoding element category (boxplots of various colors in Fig. 6). The overall nucleotide percentage of each annotation over the genome was used as the background (black dash lines in Fig. 6). In one instance representative of the large differences observed between cancer types, variants in kidney cancer genomes were found to be preferentially located in the TF binding site while lung adenocarcinoma is mutation depleted in this region (0.140 average vs. 0.098 average, in Fig. 6). A large sample difference was also observed in several cancer types. For instance, within Pilocytic Astrocytoma, there are samples that have a TF binding peak mutation fraction as high as 0.252 and as low as 0.011, which represents a ~23-fold

difference. Hence, it is important to understand the mutation patterns in these noncoding annotations, and take their unique characteristics into consideration.

Fig 6: Boxplots illustrating the distribution of variants intersecting several prominent classes of noncoding annotations in various cancer types. The percentage of mutations varies widely between noncoding element types, between cancer types, and between samples of the same cancer type.



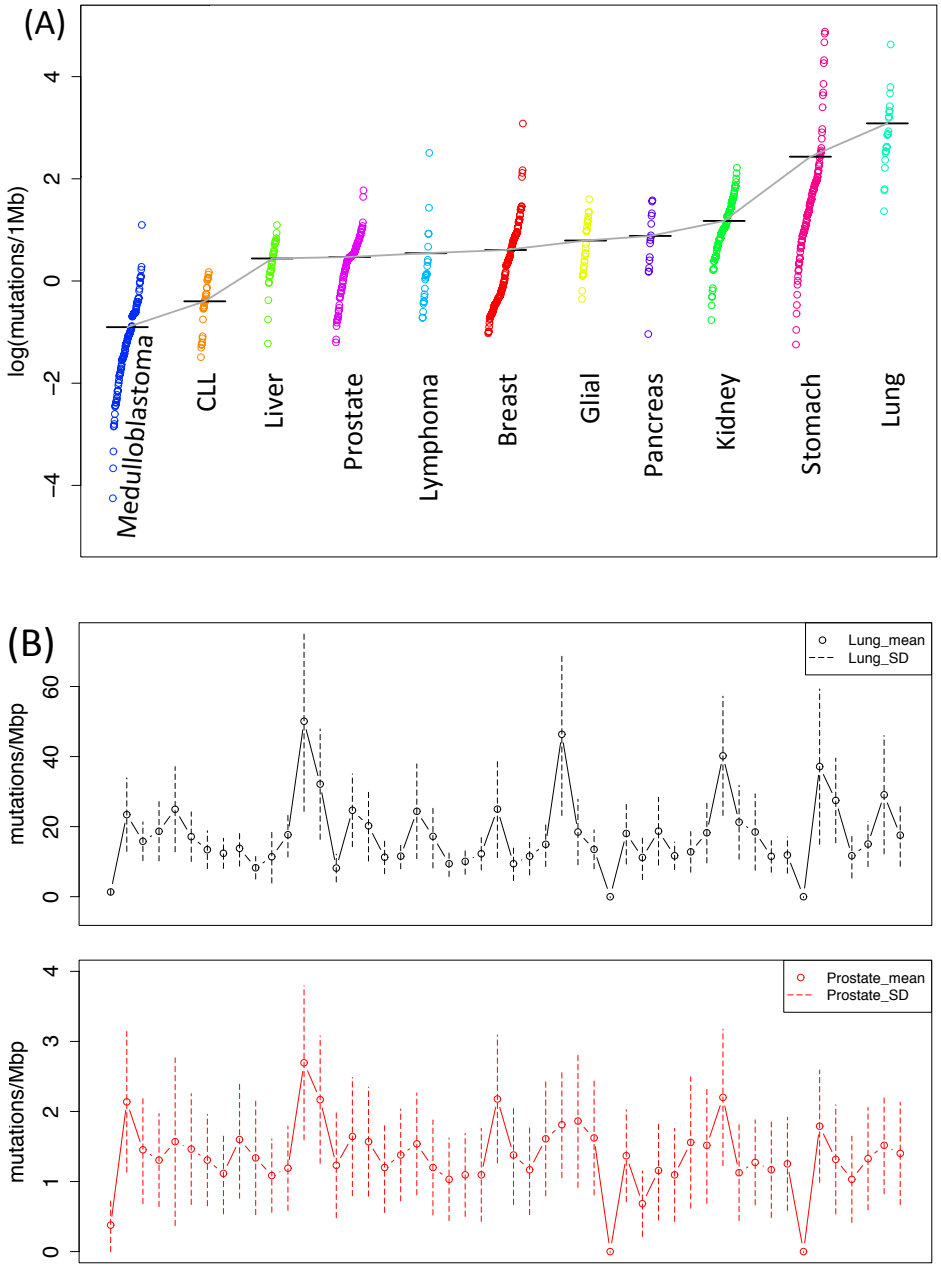
7.2. Large cancer type, sample, regional heterogeneity of cancer genomes, and the potential dependency among neighboring regions violate the binomial assumption

In Weinhold *et al.* (2014), the mutation burden tests are performed based on the binomial distribution, which inherently assumes a constant mutation rate and completely independent mutation events. However, these assumptions might not be appropriate for either somatic or germline variant analysis.

First, in our analysis of hundreds of WGS somatic mutation signatures, we observed huge cancer type, sample, and regional somatic mutation rate heterogeneity. To demonstrate cancer type and sample mutation rate heterogeneity, we selected all cancer types with more than 20 samples in it. We split the human genome into 1 mega-basepair (Mbp) size bins, and intersected the individual sample variants from our dataset to calculate the mutation rate of each sample. Consistent with the analysis in coding regions (Lawrence 2013), we observed huge mutation rate differences between cancer types. For instance, the average whole genome mutation rate in stomach cancer is as high as 11.389 mutations/Mbp (Fig 7A), which is ~800 times the mutation rate in medulloblastoma (0.0142, Fig 7A). Furthermore, the whole genome mutation rate also fluctuates wildly across samples, and such changes may go up to 100 times within the same cancer type (0.359 vs. 21.8 in breast cancer for example). Additionally, to illustrate regional mutation rate heterogeneity, we randomly selected 50 one-megabase-length regions to calculate the mean and standard deviation (SD) of the local mutation rate across samples in lung cancer and prostate cancer (Fig 7B). As shown in Fig 7B, the average local mutation rate may vary from 0 to 50.8 mutations/Mbp across the randomly selected

bins, and the SD range is unusually huge for each bin. Similar results were also observed in prostate cancer (Fig 7B).

Fig 7: (A) Between samples of the same cancer type, there is huge mutation rate heterogeneity. For most cancers, the mutation rate spans several orders of magnitude. (B) Variation in the mutation rate across chromosome 1 in lung cancer (top) and prostate cancer (bottom).



Several biological signatures could partially explain the observed mutation rate heterogeneity. For example, the later replicating regions usually suffer from

accumulative DNA damage, and therefore are prone to mutations (Stamatoyannopoulos 2009). Furthermore, methylated cytosines in CpG sites are often unstable and undergo deamination to thymine, which yields a C to T transition (Hodgkinson 2011). Hence, there is a noticeable mutation rate difference at CpG and non CpG sites. Several other hypotheses were also proposed and summarized in Hodgkinson and Eyre-Walker's review paper (2011).

Second, mutation events might not be independent of each other. For example, in germline mutation analysis, mutations with high LD are prone to co-occur. Additionally, some passenger mutations are generated by other driver mutations. The driver mutation might be a mutation in a DNA replication or repair gene. Moreover, some structural variations, such as long insertions or deletions, might cause problems in pairing during meiosis and thus generate additional point mutations in neighboring regions (Tian 2008). Consistent with this hypothesis, the mutation rates of the surrounding structural variations are elevated in several eukaryotic species (Tian 2008, Hollister 2010, McDonald 2011).

Perhaps due to the violation of these two assumptions, we observed a much higher than expected variance in the mutation count data. For example, at a 10kb bin resolution, the observed mutation count variance is 7.679 times the expected value under the binomial assumption. Hence, it is necessary to introduce other statistical models to handle such overdispersion in the mutation count data.

7.3. Improved mutation count fitting through a beta-binomial distribution

As discussed in the previous section, a binomial distribution model used in Weinhold *et al.* (2014), which assumes a constant mutation rate and independent

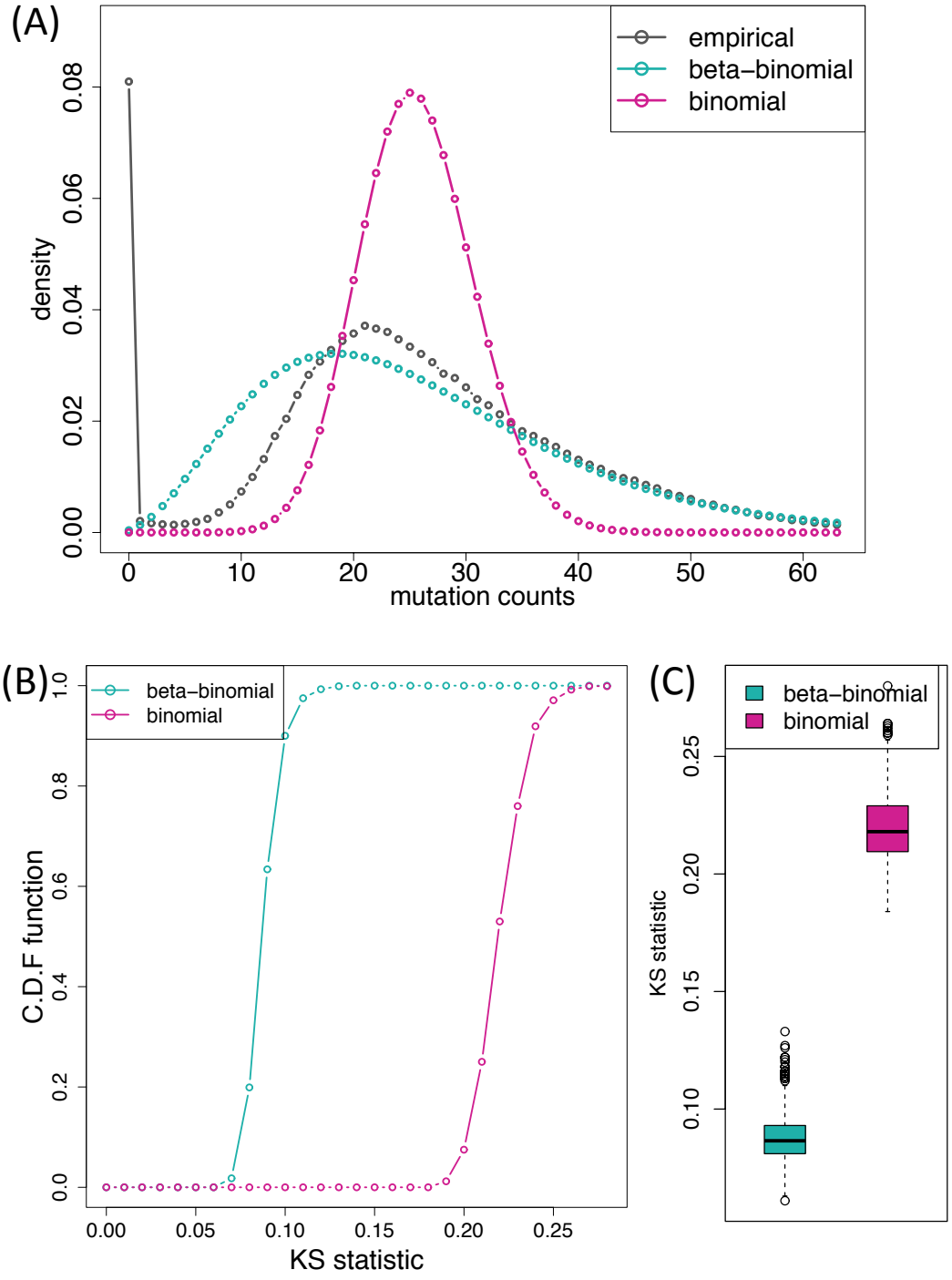
mutation process, could be problematic in more practical data analysis applications when the mutation counts are highly overdispersed. Hence, we first proposed a two-layer model to fit the variant count data (model 2 in Section 6.4). Instead of setting a constant mutation rate, our model treats the mutation rate as a beta-distributed random variable, which flexibly provides the underlying mutation rate with desired mean and variance properties. Then the mutation counts within each regulatory element could be easily modelled as a beta-binomial distribution (further details in Section 6.4).

We fitted the mutation count data at a 10kb bin resolution of the 760 WGS cancer genomes under the fixed (binomial) and variable (beta-binomial) mutation rate assumptions in Fig. 8. We calculated the frequency of the observed mutation count in each bin and compared it with the binomial (model 1) and beta-binomial (model 2) fittings respectively. It is shown in Fig. 8A that the observed data demonstrates much heavier tails than the binomial distribution, while the beta-binomial distribution fits the right tail very well. In order to quantitatively exhibit the improved performance of beta-binomial fitting, we utilized Kolmogorov-Smirnov (KS) statistics to compare the two distributions with the observed data in a nonparametric way. A larger KS statistic indicates a higher level of deviation between the two distributions. Specifically, 1000 bins were drawn from beta-binomial and binomial fitted distributions separately to calculate the KS statistic against the randomly sampled 1000 mutation counts from the observed data. This scheme was repeated 1000 times and the cumulative distribution function (C.D.F) of the KS statistics were given in Fig. 8B. The median KS statistic value for the beta-

binomial distribution was 0.087, significantly smaller than 0.218 of the binomial distribution (p-value for two-sided Wilcoxon test < 2.2×10^{-16}, boxplots given in Fig. 8C). Different bin sizes were analyzed using the sample method and results were similar (100kb bins in Fig. S3, 1kb bins in Fig. S4). In order to avoid overfitting, we utilized half of the data for distribution fitting, and the remaining half as the input to calculate the KS statistic for evaluation. This scheme was repeated 100 times. The beta-binomial distribution still significantly outperforms the binomial distribution (0.0821 vs. 0.216, p-value for two sided Wilcoxon test < 2.2×10^{-16}, Fig. S5). Hence, the improved performance of the beta-binomial distribution is due to its enhanced flexibility to handle the overdispersed mutation count data instead of overfitting.

In the significance analysis, p-values were usually calculated from the right tail of the null distribution. However, the huge deviation of the binomial distribution from the observed one could potentially introduce huge p-value inflation, and consequently result in numerous false positives. We defined the p-values for the observed distribution as the percentage of bins with equal or larger mutation counts. However, the improved fitting of the beta-binomial distribution could solve this problem and provide more accurate p-value assessment.

Fig 8: (A) The beta-binomial distribution (pink line) provides better fitting to the observed mutation counts at 10kb resolutation (black line) of 760 cancer genomes, especially at the right tail as compared to the binomial distribution (turquoise line). (B) A comparison of the cumulative distribution function (CDF) of the binomial distribution and the beta-binomial distribution from part A. (C) Boxplots of the Kolmogorov-Smirnov (KS) statistics for the two distributions.



7.4. Local background mutation rate calculation through replication timing correction further controls false positives and false negatives

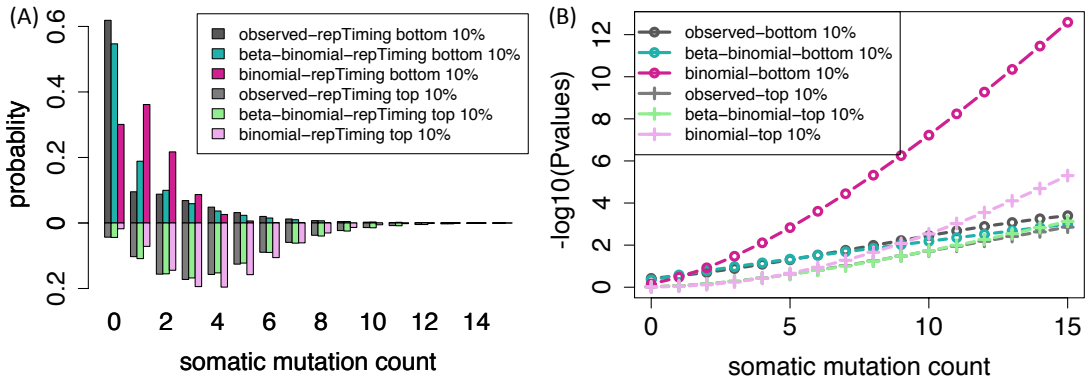
Recently, several computational efforts have been made to link somatic mutation rates with several genomic features in protein-coding regions (Hodgkinson 2011, Lawrence 2013). A particularly well-known example is DNA replication timing. During replication, the single stranded DNA usually accrues endogenous DNA damage, such as oxidation and deamination (Stamatoyannopoulos 2009). Hence, DNA that is replicated in a later stage would be susceptible to the effects of accumulative damage, and would be prone to all classes of substitutions. Consistent with this assumption, scientists observed that the later replicating regions demonstrate remarkably higher mutation rates (Stamatoyannopoulos 2009). Although replication timing has been used successfully to calculate the background model in the coding regions, little work has been done in the noncoding regions in cancer genomics. Hence, we explored the effect of replication timing on the mutation rate calculation (model 3 in Section 6.4), and the consequential effect on the p-value evaluation.

Using 1kb bins, we counted the average replication timing value within each bin, and then separated the top and bottom 10% of replication timing bins for mutation rate calculation. As shown in Fig. 9A, we observed noticeable differences in the mutation rate vis-a-vis the replication timing signal. The average mutation count of the 760 samples was 1.200 for the bottom 10% replicating timing bins, as compared to 4.028 for the top 10% (p-value for two-sided Wilcoxon test < 2.2×10^{-16}). A KS test was performed to determine whether these two sets of

mutation counts data follow the same distribution, and the p-value is less than 2.2×10^{-16}, indicating that the two distributions are significantly different.

Moreover, we observed that the mutation counts data for bins with similar replication timing values still shows extensive overdispersion. For example, for the bottom 10% of replication timing bins, the observed variance of mutation counts was 4.168, which is 3.477 times that under the binomial assumption. Consistently, we observed poor fitting of the binomial distribution against the observed distribution, especially in the right tails (Fig. 9A). The huge deviation in the right tails would result in huge p-value calculation inflation as shown in Fig. 9B. The p-value for 16 mutations in the bottom replication timing 1kb region from the empirical distribution shows only marginal significance (3.994×10^{-4}), but the binomial distribution could inflate it to 2.585×10^{-13} due to its bad fitting of the heavy tails on the right side. But our beta-binomial distribution rigorously controls the p-values through the flexible mutation rate assumption (p-value = 1.002×10^{-3}). We demonstrated the better p-value curve of the beta-binomial distribution in a variety of data points and replication timings, indicating the robustness of our method (Fig. 9B).

Fig 9: (A) The 1 kb genome bins representing the top 10% and bottom 10% of the DNA replication timing were used to derive an observed distribution of mutation counts, demonstrating the influence of replication timing. The fitted binomial and beta-binomial distributions are plotted as bar plots. (B) P-values at different mutation counts were given by the observed, beta-binomial, and binomial distribution. The binomial distribution's p-values demonstrate an inflation that is not observed in either the observed or beta-binomial distribution's p-values.



Additionally, the replication timing effect correction further improves the p-value calculation to avoid potential false positives and false negatives. For instance, for a region among the top replication timing regions, 8 mutations in 1kb bin would give a p-value of 0.094 after replication timing correction from the beta-binomial model, but might be reported as positive when ignoring replication timing effect (p-value = 0.038 from beta-binomial by mixing the top and bottom 10% replication timing points). Similarly, a p-value of 0.064 would reject 7 mutations within 1kb bin as significant without correction. However, if this point comes from the bottom 10% of replication timing regions, the true p-value should be 0.030 due to its relatively lower local mutation rate. Hence, it is important to perform covariate correction before calculating p-values.

7.5. LARVA discovered a list of highly recurrent noncoding regulatory regions from WGS data

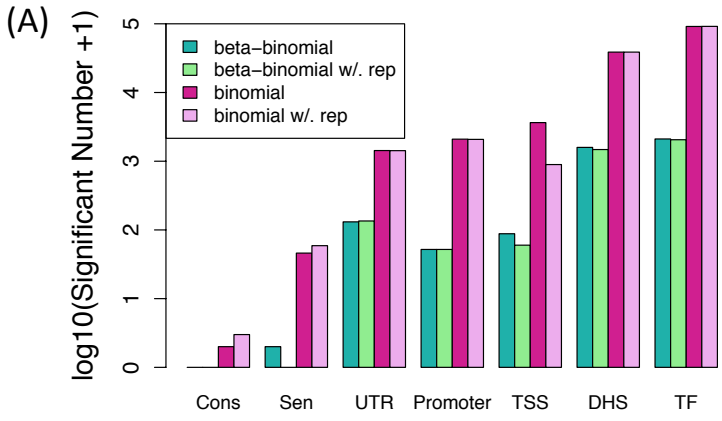
We first applied LARVA to the 760 genomes' variants, intersecting them with the noncoding regions listed in Table 7. In total, LARVA reported 3964 and 3776 highly mutated regions before and after replication timing corrections, respectively (as shown in Table 8). On the other hand, the binomial distribution models reported at least 30 times more regions as significant because of the aforementioned p-value inflation, giving rise to a high false positive rate. We also tested the immediate 100bp upstream of every possible transcription start site (see Section 6.3 for details), the results of which are depicted in Fig. 10B. Forty-five TSSs passed the 0.05 p-value thresholds after p-value adjustment (BH method, Benjamini & Hochberg 1995). Consistent with previous studies, we observed that the TSS for TERT came up in the top regions (Fig. 10B), and the oncogene TP53 also ranked second among all sites. LMO3, which ranked third after replication timing correction, is a protein coding oncogene that is predominantly expressed in brain tissue. It has been reported to be involved in a variety of cancer types, such as lung cancer (Kwon 2012) and neuroblastoma (Isogai 2011). PRRC2B's TSS was reported as the most significantly recurrent region among all TSSes. It is a protein coding gene that is extensively expressed in brain tissue, but to our best of knowledge, there is no study to show the link of PRRC2B to cancer. Further investigations should be performed for the purpose of validation. Similar results were given for promoters and UTR regions as well. We selected all the genes with highly mutated TSSes, promoters, or UTRs (adjusted p-values after corrections ≤ 0.05) and performed GO analysis

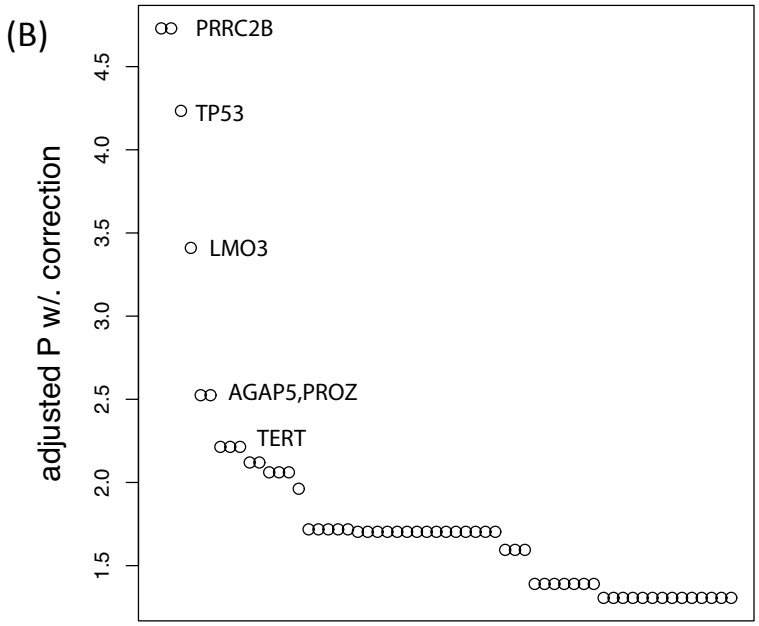
(Ashburner 2000). The top three enriched GO terms are: “negative regulation of fibroblast proliferation”, “regulation of extrinsic apoptotic signaling pathway in absence of ligand”, and “regulation of cell growth”.

Table 8: The number of highly recurrent regions in each class of noncoding annotation discovered by LARVA and the binomial distribution-based model, with or without replication timing (RT) correction.

	LARVA without RT correction	**LARVA with RT correction**	**Binomial without RT correction**	**Binomial with RT correction**
Ultra-sensitive sites	1	0	47	66
DRM Enhancers	9	6	181	183
UTRs	169	164	1415	1402
Promoters	47	47	2277	2264
TSSs	45	45	3835	932
DHS Sites	1605	1491	40316	40311
TFBSs	2112	2063	88609	88693
Sum	3988	3816	136680	133851

Fig 10: (A) The number of significant p-values implied by beta-binomial distribution and binomial distribution (with and without DNA replication timing correction). (B) A sorted p-value plot of the top significant TSSes derived from the LARVA analysis.





In terms of transcription factor binding sites, LARVA claimed 2054 out of the 5,710,954 tested binding sites are highly recurrent (0.036%). The transcription factor CTCF had 852 binding sites reported as significant (Table 9). CTCF is a multifunctional protein that is linked with multiple cancer types (Filippova 2008). Specifically, several studies have reported that disruption of CTCF binding sites

through mutations or abnormal methylation sites is closely associated with cancer (Ohlsson 2001, Takai 2001). Moreover, we found that the oncogene BCL3 has a noticeably higher significant percentage with respect to the average (7.721 times of the average, p-value for two-sided binomial test = 6.762×10^{-13}). Interestingly, BCL3 is a proto-oncogene candidate which is closely associated with progression of diverse solid tumors (Maldonado 2011). For example, BCL3 is aberrantly up- and down-regulated in breast cancer and nasopharyngeal carcinoma, respectively, and is also reported to be strongly associated with survival in colorectal cancer. However, it is not a highly mutated gene according to our data: BCL3's mutation rate is 1.22 mutations/Mbp while the gene average is 2.52 mutations/Mbp. Our analysis suggests another possibility—that the misregulation of BCL3 is possibly due to binding site disruption instead of the changes in the protein itself. Further computational and experimental effort should be made to clarify the mechanism of BCL3 regulation in different cancer types.

Table 9: A summary of the highly recurrent transcription factor binding sites (TFBSs).

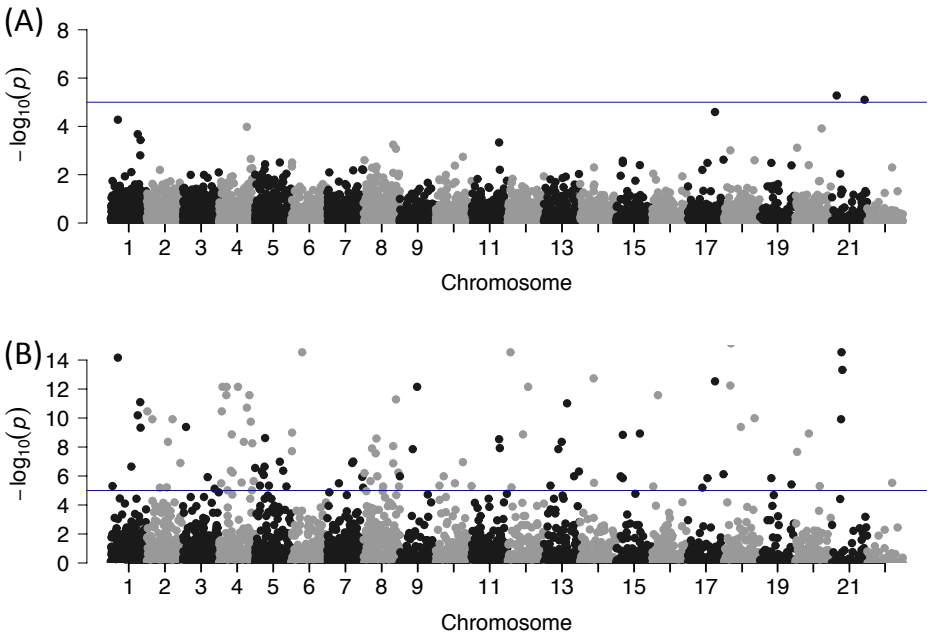
TF	Significant	Total	Percent
CTCF	870	2659116	0.000327176
RAD21	133	352066	0.00037777
MAFK	90	112696	0.000798609
CEBPB	77	111583	0.000690069
SPI1	53	81053	0.000653893
STAT3	46	140549	0.000327288
NR2C2	44	4555	0.009659715
MYC	39	129206	0.000301844
NFKB1	39	85426	0.000456535
SMC3	31	73788	0.000420123
MAX	30	84339	0.000355707
JUND	29	75579	0.000383704
BCL3	22	7806	0.002818345
EP300	22	87290	0.000252033
FOXA1	22	88981	0.000247244
USF1	22	70983	0.000309933
GATA2	20	70723	0.000282793

7.6. Whole genome recurrent events evaluation

Despite great efforts to annotate noncoding regions, there are still many regions with as yet unknown regulatory roles. In order to evaluate the recurrent events in these regions, LARVA provides all possible p-values, whether before or after adjustment, and with or without replication timing corrections, for high confidence bins on the genome (see Section 6.5 for details) of variable length. We also compared the results from our beta-binomial model with the binomial model. For example, we randomly sampled five thousand 10kb bins from the whole genome and made a Manhattan plot of p-values from both methods. It is obvious that the p-values from the binomial distribution were noticeably inflated (Fig. 11B), while our beta-binomial model effectively controls the p-values (Fig. 11A). Consistent with this result, we found that p-values from LARVA follow a uniform distribution much better than those from binomial distribution (Fig. S13). We want

to emphasize that as the sample size grows larger (such as in the following section, which describes our LARVA exome analysis), and as the target region grows larger, we expect more severe deviation from the constant mutation rate assumption, usually resulting in better performance for LARVA compared to the binomial model.

Fig 11: Manhattan plot of the p-values from 5000 randomly samples 10kb bins from the beta-binomial distribution (A) and the binomial distribution (B). The binomial distribution may provide heavily inflated p-values due to its inadequacy to capture the extensive overdispersion of the mutation count data.



7.7 Coding region calibration

It is difficult to rigorously test LARVA's sensitivity and specificity due to the lack of a benchmark dataset. In contrast to our expectations for the coding regions, we have less information for how LARVA should behave on noncoding regions. Thus, although LARVA is not optimized on coding region analysis, we re-estimated the background model on just the coding regions. In particular, given our better understanding of coding cancer drivers, we have evaluated LARVA on coding

regions on a total of 5,032 whole exome sequencing samples from TCGA (see Supplement for details). To compare the beta-binomial model with the binomial model we used a consistent and conservative threshold for both.

Many highly mutated genes discovered by LARVA were clearly documented as associated with some type of cancer. On the other hand, many false positives were reported by the simple binomial test. Moreover, p-values calculated from LARVA follow a uniform distribution quite well, and our replication timing correction further improves the p-value distribution (Fig. S12). However, the p-value distribution from the binomial model severely violates the uniform distribution assumption, providing further evidence of the binomial model's inappropriate fitting.

8. Complementary Computational Tools

8.1. ACT

In addition to LARVA, there are a number of other useful computational analyses that would be useful to conduct with whole genome data. We have implemented a number of these workflows in our Aggregation and Correlation Toolbox (ACT), an efficient, multifaceted toolbox for analyzing continuous signal and discrete region tracks from high-throughput genomic experiments, such as RNA-seq or ChIP-chip signal profiles from the ENCODE and modENCODE projects, or lists of single nucleotide polymorphisms from the 1000 Genomes Project. ACT is able to generate aggregate profiles of a given track around a set of specified anchor points, such as transcription start sites. It is also able to correlate related tracks and analyze them for saturation—i.e. how much of a certain feature is covered with each new succeeding experiment. The ACT site contains downloadable code in a variety of formats, interactive web servers (for use on small quantities of data), example datasets, documentation and a gallery of outputs. This section explains the components of the toolbox in more detail, and its applications in various contexts.

There is now an abundance of genome-sized data from high-throughput genomic experiments. For instance, there are ChIP-chip, ChIP-seq and RNA-seq experiments from the ENCODE (Birney 2007) and modENCODE (Celniker 2009) projects. There are also genome sequence data that can be used to generate tracks measuring sequence content, such as the densities of single nucleotide polymorphisms (SNPs) from dbSNP (Sherry 2001) and the 1000 Genomes Project (Durbin 2010). In most cases, the representations of these data take the form of

either signal tracks that describe a genomic landscape or distinct region tracks that tag portions of the genome as active. The aggregation and correlation toolbox (ACT) provides a powerful set of programs that can be applied to any experiments producing data in these formats. The ability to analyze multiple genomic datasets is important, as demonstrated by tools like Galaxy (Giardine 2005). ACT provides a unique set of functionality that complements existing methods of analysis.

8.1.1. ACT Overview

ACT facilitates three main types of analysis:

Aggregation: In many scenarios, it is useful to determine the distribution of signals in a signal track relative to certain genomic anchors (Fig. 12, aggregation). For example, it has recently been reported that the contribution of each transcription factor binding site to tissue-specific gene expression depends on its position relative to the transcription start site (TSS) (MacIsaac 2010). It is thus useful to aggregate binding signals of transcription factors at a certain distance from the TSSs of all genes (the anchors). In general, this type of aggregation analyses helps identify proximity correlations and functional relationships between the signals and anchors. In the ENCODE pilot study (Birney 2007), aggregation analysis was used to demonstrate positional relationships between chromatin features and TSSs.

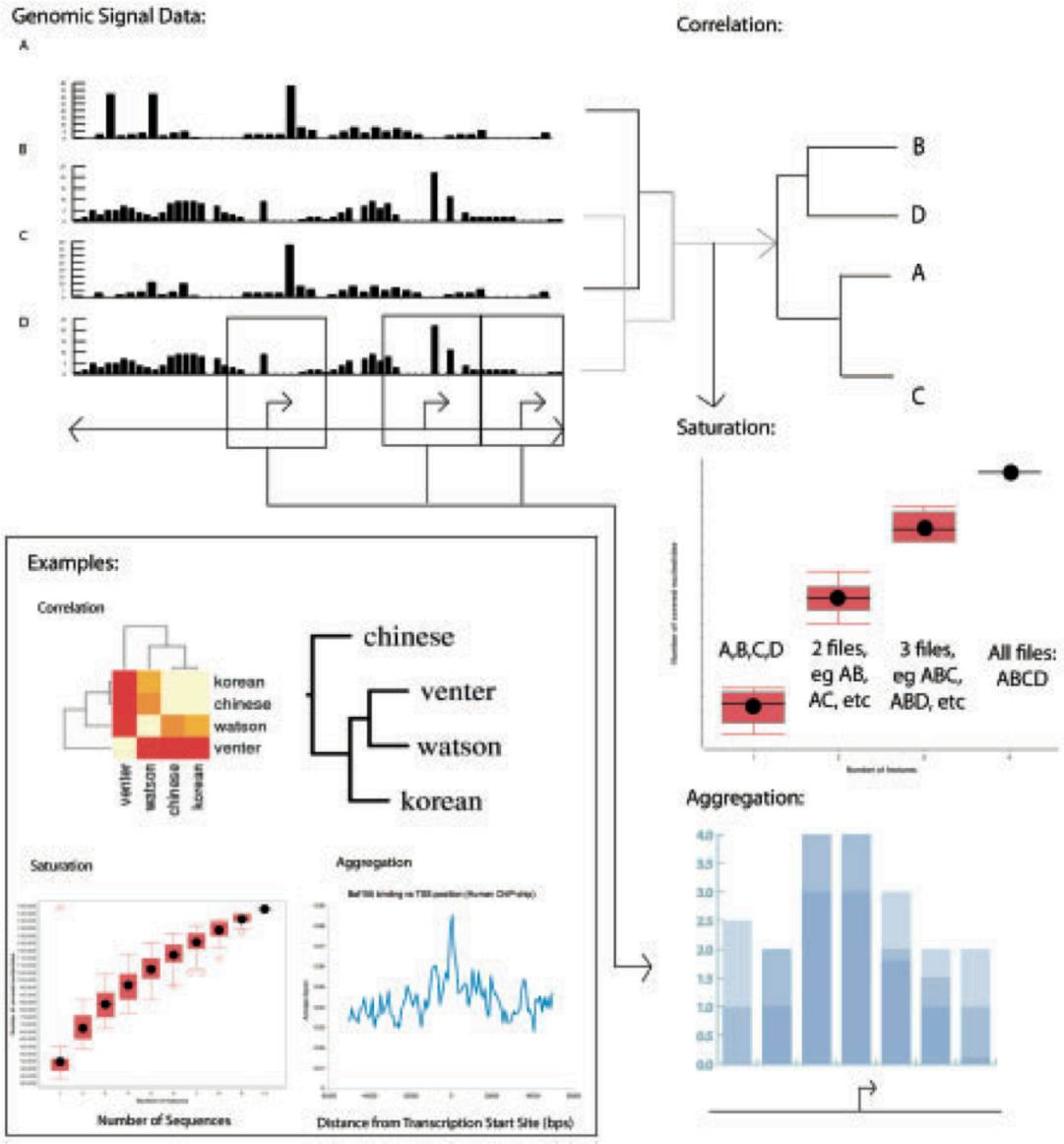
Correlation: It is also useful to consider how multiple related signal tracks are correlated with each other. For example, a previous study (Zhang 2007) demonstrated, using whole-track correlation methods, that there was a consistent relationship among transcription factors as judged by their signal profiles across several ChIP-chip experiments. By providing a means of correlating signal tracks

with each other, ACT allows for initial comparison of different experiments to see which are more similar or related than others (Fig. 12, correlation).

Saturation: Another important type of analysis is determining the number of experimental conditions required to achieve a high genomic coverage of the biological phenomenon under study. For example, using ChIP-chip or ChIP-seq experiments, one could identify a set of transcription factor binding sites from a human cell line. When the experiment is repeated using another cell line, some additional binding sites could be identified. How many cell lines need to be considered in order to reach the point of saturation, so that few new binding sites would be identified by extra experiments? ACT produces plots that help answer this type of question.

ACT was developed primarily by Justin Jee, with supporting contributions from myself and others (Jee 2011). My specific contributions included the development of portions of the saturation tool, as well as improvements to the efficiency of the implementation of the correlation tool.

Fig. 12: Uses of ACT using signal tracks from various sources. Signal around all TSSs is aggregated to give an average signal profile, for example of Baf155 binding around TSSs (Encode Project) (aggregation). Figure made in Excel (correlation). Multiple signal tracks are correlated to show which tracks are more or less related to each other. In the selected example, a heatmap of the SNP track correlation between four individuals (dbSNP) leads to a dendogram of their phylogenetic relationship. Figure made using Web ACT. Each additional signal track increases the number of base pairs covered (saturation). When the addition of signal tracks is considered in all possible combinations, the average increase in coverage, with error bars, can be visualized by a saturation plot. In the example, data are taken from individuals from dbSNP [with additional genomes from Ahn *et al.* (2009), Bentley *et al.* (2008), Drmanac *et al.* (2010), Kim *et al.* (2009)]. In each box plot, the top and bottom pink bars correspond to the maximum and minimum normal values, the top edge, middle line and bottom edge of the box correspond to the top 25 percentile, median and bottom 25 percentile, the black dot is the mean, and red circles are outliers. Figure made using ACT downloadable saturation program.



8.1.2. Details and Use Cases

ACT is available as a suite of downloadable scripts corresponding to the aggregation, correlation and saturation components of the toolbox. The tool is intended for Linux/Unix users with Java and Python. In addition, it is useful to have R for output visualization for the aggregation and correlation tools. There is also a compendium of other versions of the tool components written in different languages and with varied functionality. For some types of analysis, there are web components for demonstration purposes on small datasets with built-in visualization features. However, because most whole-genome signal tracks are too large to upload via standard Internet connections, users are recommended to download the toolbox and run it locally. As performing these calculations on whole-genome data can be especially time intensive, the version of the tools presented here has been designed to run efficiently on large datasets.

Aggregation: The aggregation component is designed to take a signal track (.sgr or .wig) and an annotation track (.bed) as input, and compute the average signal over a certain number of base pairs upstream and downstream of (i.e. a fixed radius around) the annotations. In other words, signal values are taken from the region surrounding each annotation, and averaged over the number of annotation anchors provided. The base pair resolution of the aggregation can be specified by the number of bins (narrower bins give more data points and therefore finer granularity). Results of such calculation can be plotted as in Fig. 12 (aggregation). ACT also provides features such as computing the standard deviation, median and quartiles that can be viewed as a boxplot, as well as scaling aggregation over regions

such as areas between transcription start and end sites or within exons so that all of the aggregate signals within those regions fall into a fixed number of bins. In this case, bin size is dynamically computed for each region so that the same number of bins cover regions of different sizes.

Correlation: The correlation analysis takes a set of active genomic regions (.bed) such as a SNP track or a genomic signal track (.wig). It then divides genomic coordinates into bins and gives each bin a value corresponding to the mean or maximum signal values which fall within the bin, or assigns value based on the number of 'active regions' which fall within the bin. A final correlation matrix is created based on either the Spearman's, Pearson's or normal score correlation between each pair of binned datasets. The results can be visualized as a heatmap or as a phylogenetic tree using programs such as PHYLIP (Felsenstein, 1996). One version of the correlation tool uses parallelization to decrease the program's overall running time. This component was written largely in Java. Examples of correlation output based on SNP tracks and ChIP-chip data are shown in Fig. 12 (correlation).

Saturation: We provide an efficiently implemented saturation plot generator. Each input file corresponds to one dataset (e.g. one new individual, in .bed format), and each line in a file specifies a genomic location that has the biological phenomenon under study (e.g. tagged SNPs). The saturation plot shows, with each new dataset (*x*-axis), what percentage of genomic base pairs are covered (*y*-axis). The program considers the various combinations in which tracks can be added so that the increase in base pair coverage is a range of values based on all the files in the input. The resulting plot is output in PDF format (Fig. 12, saturation), in which a

series of boxplots depicts increasing base pair coverage, where the boxplot at each position *m* on the *x*-axis shows the coverage values of all combinations of *m* conditions. Boxplots that approach a horizontal asymptote indicate that the coverage has reached saturation. Our implementation makes use of special data structures to avoid redundant counting. It normally takes less than a minute to generate the plot for up to 30 input files each with a few thousand lines. To handle more files and files with more lines, the tool also provides an option to compute the coverage of a random sample of the input file combinations.

8.1.3. Discussion

There are number of additional analyses that can be done to finetune the output of ACT. For instance, it is possible to use the online genomic signal aggregator (GSA), which assigns each genomic position to the nearest anchor in order to reduce the artifacts caused by the subsets of anchors clustering together, to handle tightly clustered anchors. Also, aggregation can be used in conjunction with genome structure correction to determine if the enrichments of a given signal with respect to anchor points are significant relative to the non-random positioning of the anchors (Birney 2007). This correction takes into account the fact that a 'random' distribution of anchors on the genome arises from a distinctly non-uniform distribution. Practically, this could be carried out through ACT by comparing the aggregation over anchors (e.g. TSSs) to that from 'randomized anchors', where the latter is generated by shifting anchor coordinates along the chromosome or transferring anchor coordinates from a second chromosome to the one of interest.

Finally, ACT can be used as a starting point for other downstream analyses. In the instance of RNA-seq data tracks, further analysis can be conducted with RseqTools (Habegger 2011) to, for example, determine additional similarities between two or more highly correlated tracks. The results of correlation analysis, for instance, can also be fed into downstream principal component analysis, allowing for grouping of coregulating factors with their coregulated sites. This would simply involve diagonalization of the output correlation matrix from ACT. Saturation analysis can also be used to inform future experimental design.

9. Conclusions & Future Work

Due to the rapid decline in time and money involved to perform whole genome sequencing, data is now available for thousands of genomes where previously only a handful were available (Shendure 2008). However, the analyses necessary for finding useful patterns in this data, and making sense of it for clinical benefit, have not kept pace with this sudden increase. Therefore, it is important that new algorithms are developed that can efficiently mine relevant patterns from genome sequence data, and that user interfaces for finding and understanding that data are optimized so that clinicians and biologists, who may not have extensive technical expertise, can use these results effectively in their work.

Compared with the extensive computational and experimental efforts on the mutation patterns in the protein coding regions in the past decade (Koch 2014), the noncoding regions, which were viewed as "dark matter", and comprise up to 98% of the human genome, are less investigated in cancer research studies, partially due to limited knowledge of noncoding function. However, recently several examples clearly pinpointed the phenotypic effect of mutations in noncoding regulatory regions in a variety of cancer types. For instance, the TERT promoter, a well-known example, has been associated with several cancer types (Grossman 2013, Maurano 2012, Vinagre 2013). Fusions of the 5' UTR of TMPRSS2 with ETS genes frequently observed in prostate cancer, as well as mutations in certain miRNA binding sites (Lin 2012), can influence the binding affinity at these sites, and thus affect androgen receptor regulation in prostate cancer. Hence, it is important to explore the mutation landscapes of such noncoding regions.

In this dissertation, we have introduced a new computational framework for exploring patterns of mutation in the noncoding regulatory regions of human genomes. Unlike coding region analyses, where burden tests may be conducted with naturally defined segments—genes—and synonymous sites may serve as a biologically meaningful background, whole genome burden tests are hindered by the fact that many noncoding functional regions are poorly defined, if at all. We took advantage of the complete genome annotation efforts of the ENCODE project (Dunham 2012) to extract the most extensive catalog of noncoding regulatory regions to date. We included the TF binding sites and DHS sites from all ENCODE experiments, promoters, UTRs, predicted enhancers, conserved and sensitive noncoding regions from our previous efforts (Fu 2014). These annotations are tested for mutation burden, and the functional significance of each highly mutated region is immediately clear. Hence, LARVA's complete design, in terms of both software and provided data, offers a new, convenient processing engine for whole genome mutation burden analysis.

We then ran our algorithm on 760 cancer genomes using the comprehensive list of noncoding annotations to search for highly mutated regulatory regions as potential noncoding driver candidates. Consistent with the highly heterogeneous protein coding regions (Lawrence 2013), we observed larger than expected mutation variation across cancer types, samples, and genomic regions (Fig. 7). Therefore, the recently proposed binomial models, which assume a constant mutation rate and independence of mutation events, might be inadequate for the observed data (Fig. 8, Fig. S3-S4). Instead, we set up two hierarchical models to

handle mutation count overdispersion (model 2 and model 3 in Section 6.4). First, we flexibly modeled the mutation rate in the regulatory elements as a two-parameter beta distribution $beta(\mu,\delta)$, resulting in a beta-binomial distribution for the variant counts; $beta(\mu,\delta)$ can be seen as the distribution from which the whole genome region-specific mutation rates (p) are sampled. Alternatively, $beta(\mu,\delta)$ can be treated as the distribution from which patient-specific or cancer type-specific mutation rates are sampled. Therefore, when analyzing large regions, such as enhancers that might be over 10kb, or small regions (such as 200bp TSS sites) in cohorts with a large number of samples, the beta-binomial model provides improved fitting over the binomial model. On the other hand, when the target region is small, or the patients are more homogeneous, we expect less overdispersion from the data. Then, the estimated beta-binomial parameters will be similar to those of the binomial distribution.

In addition, genomic features, such as replication timing, expression level, and GC content, have a major effect on the background mutation rate (Fig. S6)(Lawrence 2013). As a consequence, the overall background mutation rate is actually a mixture of several different distributions, resulting in extra variance in the mutation count data (Fig. S14). Therefore, it is necessary to separate the covariate effects. In this dissertation, we found replication timing is the feature that explains the largest amount of variation in the mutation counts data, so we started from this major covariate and corrected its effect by estimating the local mutation parameters in the beta-binomial model. In the future, we plan to further correct multiple covariates jointly. Moreover, in general the quality of LARVA output depends on the

quality of the input variants. There are some known artifacts in the earlier variant call sets which might introduce biased results. In the future, the release of large scale uniformly processed variant call sets will definitely improve subsequent LARVA analyses.

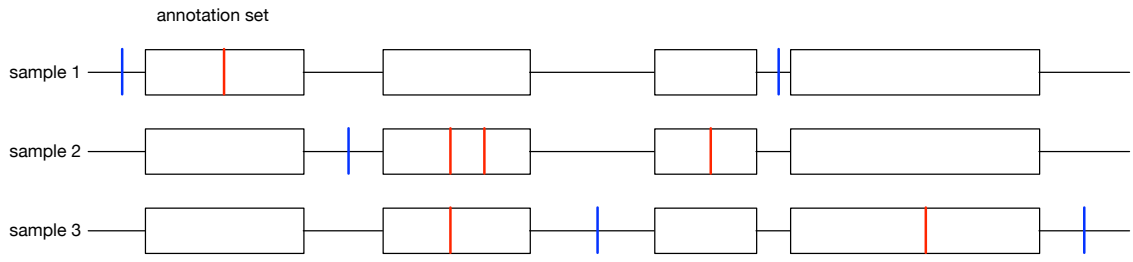
In the 760 cancer whole genomes in our analysis, we discovered 3776 noncoding regulatory regions that have significantly higher mutations than expected and provided the mutation enrichment significance of bins with variable length on the whole genome (Table 8). A list of known noncoding hypomutated regions, such as TERT and TP53 TSS, were also reported by our analysis, which convincingly proved the effectiveness of LARVA in discovering functionally relevant results. We also observed some relatively novel results such as PRRC2B TSS, CTCF and BCL3 binding sites. BCL3 is a known oncogene that is highly associated with several solid tumors (Kim 2008, Maldonado 2011), but this gene itself is not enriched in our analysis. Our results advocate an alternate possibility: its mutation in cancer cells is actually in the disruption of its binding sites, rather than the disabling of the protein itself. We released our annotations to the public, which would potentially serve as a useful resource for cancer researchers in the future.

It is worth pointing out that although LARVA was designed to analyze somatic variants, it can be immediately extended to discover the hypermutated regions for germline variants. As with somatic variants, the germline mutation landscape demonstrates extensive heterogeneity and dependency, which can't be properly handled by a binomial distribution. Furthermore, unlike GWAS common variants discovery, LARVA could combine both rare and common variants to assess

the mutation burden in noncoding regulatory regions. Due to the popularity of studying rare variants in human genomes, LARVA could potentially serve as a powerful tool to discover hypermutated noncoding regulatory regions.

LARVA's future capabilities may include the study of genetic diseases at a systems level, with pathway analyses and interaction network analyses. This use of LARVA extends the idea of finding recurrently mutated annotations to a larger scope. Instead of looking for variants from multiple samples in a single annotation, LARVA would look for the accumulation of variants from multiple samples across a group of annotations (Fig 13). These annotation groups would correspond to functionally related annotations, such as a series of metabolic pathway enzymes, or a cluster of physically interacting proteins. The application of LARVA to aggregating variants on the metabolome, using pathway databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa 2000, 2011), and on protein-protein interaction (PPI) networks, using pairwise interaction databases such as the Human Protein Reference Database (HPRD)(Prasad 2009), could enable a fresh understanding of cancer disruption on a system-wide level.

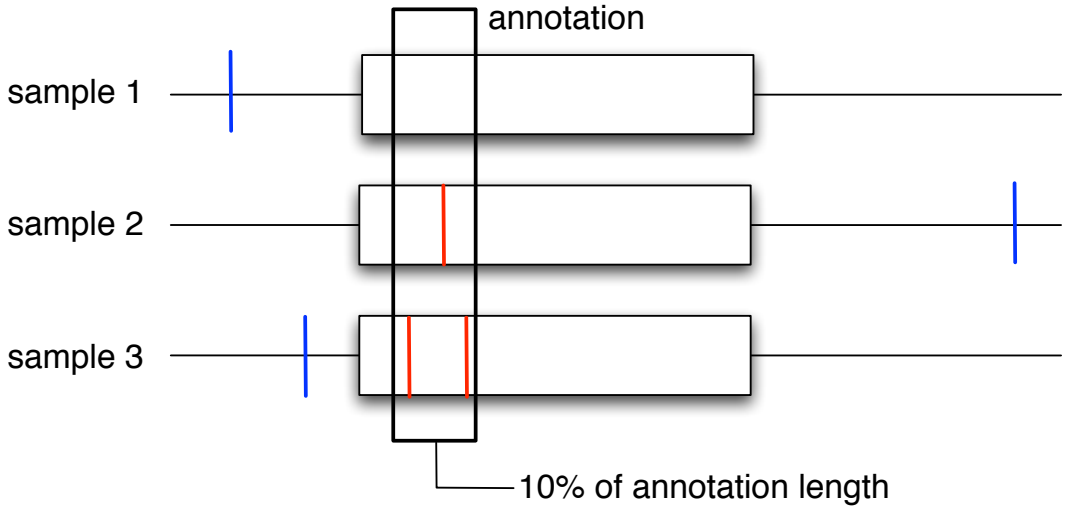
Fig. 13: A recurrently mutated annotation set. Intersecting variants (red) span multiple samples.



Another future application of LARVA would be to focus its recurrence analysis on portions of annotations. LARVA would look for significant accumulations

of variants from multiple samples in a fraction of the total annotation's length, or in specific functional domains (Fig 14). The identification of these recurrently mutated annotation domains would facilitate the discovery of the functional relevance of recurrently mutated annotations.

Fig. 14: Recurrently mutated annotation domains represent a fraction of an annotation that contains intersecting variants from multiple samples.



LARVA introduces a new model for the distribution of background mutations in cancer genomes. This model addresses a specific shortcoming of previous models that assumed a constant mutation rate over the entire genome. LARVA's background mutation model represents the mutation rate as a beta distribution, allowing the rate to vary between basepairs and regions in accordance with the observed mutation counts across a broad swath of cancer variant data. LARVA's model significantly improves the false positive rate relative to previous models, and enables the identification and prioritization of recurrently mutated annotations in an annotation set of interest.

The annotations that LARVA discovers and ranks as important can be fed into downstream analyses for further verification of the annotations' roles in cancer disruption. This includes functional annotation analyses to pinpoint the impact of mutated sites. Incorporating these features into LARVA in the future would help make LARVA more generally and broadly applicable, and bring it into a potential role as a "one-stop shop" for studying variants.

We will also continue to improve LARVA's algorithms. As the amount of genetic data increases, it will be important to further optimize LARVA's computational efficiency, and therefore we are investigating these issues for future iterations of LARVA. One particular focus for the speedup of LARVA is to produce a general purpose GPU (graphics processing unit) version of the code. GPUs are specialized processors used in computers for the specific purpose of performing the massively parallel computations necessary to render computer graphics. GPUs are also good for running other computations that can be parallelized and optimized for the onboard hardware resources of the GPU. Future research will include determining the suitability of LARVA's algorithms for GPGPU computing, and adapting the code for that type of processing.

Although LARVA may be run as a command line tool, its usage in this form requires a setup that necessitates some technical skill to accomplish, a limitation that hinders the potential reach of LARVA. The availability of a Dockerized version of LARVA mitigates the complexity of first-time setup. However, a Web version of LARVA would make it even easier for nontechnical users to take advantage of LARVA's capabilities. On the other hand, a Web version would be hampered by the

fact that the amount of data that can be used is limited by the upload and download speed of the user, making the Web version unsuitable for large scale use. There would also be the issue of properly protecting data that is under strict usage restrictions so that no unauthorized users can access the data on LARVA's servers, or in transit. These and other issues will be considered in future iterations of LARVA.

Finally, we will continue to gain insights by applying LARVA to additional cancer types and subtypes. The Cancer Genome Atlas (TCGA) is anticipated to have generated roughly 2.5 petabytes (PB) of data by its conclusion ("NCI Cancer Genomics"). Other consortia spanning many other countries and populations have similar efforts under way to study the most pressing cancers particular to those populations. With the availability of such a tremendous trove of data, LARVA's development is most opportune. LARVA's simple and efficient methods for studying recurrent mutation patterns will be invaluable for gaining new information on the molecular characteristics of cancer disruption. In the long term, we envision LARVA becoming increasingly useful for elucidating important insights and understanding about all types of genetic diseases.

In summary, LARVA is a powerful computational method to explore a broad range of genome annotations to uncover the ones that are mutated across many samples. LARVA makes it possible to predict putative noncoding drivers of genetic disease, and prioritize these predicted drivers for more rigorous downstream analysis. This may lead to faster identification of important targets that may be used to suppress disease with therapies and drugs.

References

1. Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Research* **19,** 1622–1629 (2009).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
3. Almasy, L. *et al.* Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. in *BMC proceedings* **8,** S2 (BioMed Central Ltd, 2014).
4. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).
5. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153,** 666–677 (2013).
6. Badis, G. *et al.* Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* **324,** 1720–1723 (2009).
7. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature Genetics* **44,** 685–689 (2012).
8. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).
9. Bejerano, G. Ultraconserved Elements in the Human Genome. *Science* **304,** 1321–1325 (2004).
10. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
11. Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological).* **57**, 289-300 (1995).
12. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
13. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology* **24,** 1429–1435 (2006).
14. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* (2011).
15. Berger, M. F. *et al.* Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell* **133,** 1266–1276 (2008).
16. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
17. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22,** 1790–1797 (2012).
18. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research* **21,** 456–464 (2010).
19. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research* **36,** D102–D106 (2007).

20. Cancer Genome Atlas Research *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, **45**, 1113-1120 (2013).
21. Cancer Genomics Hub - UC Santa Cruz. at <https://cghub.ucsc.edu/index.html>
22. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research* **69,** 6660 (2009).
23. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459,** 927–930 (2009).
24. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2,** 401–404 (2012).
25. Chen, C. L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research* **20,** 447–457 (2010).
26. Chen, X. *et al.* XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature* **508,** 103–107 (2014).
27. Chi, K. R. The year of sequencing. *Nature Methods* **5,** 11–14 (2008).
28. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3,** (2007).
29. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2,** 2366–2382 (2007).
30. CT, G. Primer: Sequencing—the next generation. *Nature Methods* **5,** 15 (2008).
31. D'Antonio, M., & Ciccarelli F.D. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biology* **14**:R52 (2013).
32. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Research* **22,** 1589–1598 (2012).
33. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43,** 491–498 (2011).
34. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLoS ONE* **7,** e30377 (2012).
35. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327,** 78–81 (2009).
36. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
37. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
38. Erturk, E. *et al.* Evaluation of Genetic Variations in miRNA-Binding Sites of BRCA1 and BRCA2 Genes as Risk Factors for the Development of Early-Onset and/or Familial Breast Cancer. *Asian Pacific Journal of Cancer Prevention* **15,** 8319–8324 (2014).
39. Esteller, M. Non-coding RNAs in human disease. *Nature Reviews Genetics* **12,** 861–874 (2011).

40. Filippova, G.N. Genetics and epigenetics of the multifunctional protein CTCF. *Current topics in developmental biology*, **80**, 337-360 (2008).
41. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* **15,** 480 (2014).
42. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4,** 177–183 (2004).
43. Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512,** 445–448 (2014).
44. Giardine, B. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* **15,** 1451–1455 (2005).
45. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–243 (2012).
46. Grossman, S. R. *et al.* Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell* **152,** 703–713 (2013).
47. Guo, X., Lin, M., Rockowitz, S., Lachman, H. M. & Zheng, D. Characterization of Human Pseudogene-Derived Non-Coding RNAs for Functional Potential. *PLoS ONE* **9,** e93972 (2014).
48. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27,** 281 (2011).
49. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–1774 (2012).
50. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* **12,** 756–766 (2011).
51. Hollister, J. D., Ross-Ibarra, J. & Gaut, B. S. Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants. *Molecular Biology and Evolution* **27,** 409–416 (2010).
52. Hudson (Chairperson), T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
53. Imielinski, M. *et al.* Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. *Cell* **150,** 1107–1120 (2012).
54. Isogai, E. *et al.* Oncogenic LMO3 Collaborates with HEN2 to Enhance Neuroblastoma Cell Growth through Transactivation of Mash1. *PLoS ONE* **6,** e19297 (2011).
55. Jakob, J. A. *et al.* NRAS mutation status is an independent prognostic factor in metastatic melanoma. *Cancer* **118,** 4014–4023 (2012).
56. Jee, J. *et al.* ACT: Aggregation and Correlation Toolbox for Analyses of Genome Tracks. *Bioinformatics* (2011).
57. Jiao, X. *et al.* DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28,** 1805–1806 (2012).
58. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
59. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40,** D109–D114 (2011).
60. Keshava Prasad, T. S. *et al.* Human protein reference database—2009 update. *Nucleic acids research* **37,** D767 (2009).

61. Kheradpour, P. Computational regulatory genomics: motifs, networks, and dynamics. (2012). at <http://18.7.29.232/handle/1721.1/70871>
62. Khurana, E. *et al.* Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science* **342,** 1235587–1235587 (2013).
63. Kim, J.-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* (2009). doi:10.1038/nature08211
64. Kim, Y.M. *et al.* The proto-oncogene Bcl3, induced by Tax, represses Tax-mediated transcription via p300 displacement from the human T-cell leukemia virus type 1 promoter. *Journal of virology*, **82**, 11939-11947 (2008).
65. Kleinman, J.C. Proportions with extraneous variance: two dependent samples. *Biometrics*, **31**, 737-743 (1975).
66. Koch, L. Cancer genomics: Non-coding mutations in the driver seat. *Nature reviews. Genetics*, **15**, 574-575 (2014).
67. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics* **82,** 949–958 (2008).
68. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nature Genetics* (2012). doi:10.1038/ng.2359
69. Kurtova, A. V. *et al.* Blocking PGE2-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature* (2014). doi:10.1038/nature14034
70. Kwon, Y.J. *et al.* Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, **7**, 20-33 (2012).
71. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
72. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
73. Leaf, Clifton. "Why We're Losing the War on Cancer [And How to Win It]." *Fortune* 22 Mar. 2004: 76-92. Print.
74. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Research* **39,** D28–D31 (2010).
75. Leiserson, M. D. M., Blokh, D., Sharan, R. & Raphael, B. J. Simultaneous Identification of Multiple Driver Pathways in Cancer. *PLoS Computational Biology* **9,** e1003054 (2013).
76. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).
77. Lin, P.-C. *et al.* Epigenetic Repression of miR-31 Disrupts Androgen Receptor Homeostasis and Contributes to Prostate Cancer Progression. *Cancer Research* **73,** 1232–1244 (2012).
78. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Research* gkv803 (2015). doi:10.1093/nar/gkv803

79. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences* **109,** 3879–3884 (2012).
80. Long, G. V. *et al.* Increased MAPK reactivation in early resistance to dabrafenib/trametinib combination therapy of BRAF-mutant metastatic melanoma. *Nature Communications* **5,** 5694 (2014).
81. MacIsaac, K. D. *et al.* A Quantitative Model of Transcriptional Regulation Reveals the Influence of Binding Location on Expression. *PLoS Computational Biology* **6,** e1000773 (2010).
82. Madrigal, P. & Krajewski, P. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Frontiers in Genetics* **3,** (2012).
83. Maldonado, V. and Melendez-Zajgla, J. (2011) Role of Bcl-3 in solid tumors. *Molecular cancer*, **10**, 152.
84. Mardis, E. R. ChIP-seq: welcome to the new frontier. *Nature Methods* **4,** 613–614 (2007).
85. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42,** D142–D147 (2013).
86. Matys, V. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31,** 374–378 (2003).
87. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
88. McDonald, M. J., Wang, W.-C., Huang, H.-D. & Leu, J.-Y. Clusters of Nucleotide Substitutions and Insertion/Deletion Mutations Are Associated with Repeat Sequences. *PLoS Biology* **9,** e1000622 (2011).
89. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303 (2010).
90. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
91. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
92. Medrzycki, M. *et al.* Histone H1.3 Suppresses H19 Noncoding RNA Expression and Cell Growth of Ovarian Cancer Cells. *Cancer Research* **74,** 6463–6473 (2014).
93. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
94. "NCI Cancer Genomics Cloud Pilots Concept." *NCI.* National Institutes of Health, n.d. Web. 25 Feb. 2014.
95. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993 (2012).
96. Ohlsson, R. *et al*. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in genetics : TIG*, **17**, 520-527 (2001).

97. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Computational Biology* **9,** e1003153 (2013).
98. Parmigiani, G. *et al.* Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* **93,** 17–21 (2009).
99. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol* **13,** R51 (2012).
100. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **21,** 447–455 (2010).
101. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2009).
102. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
103. Rausch, T. *et al.* Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell* **148,** 59–71 (2012).
104. Ribas, A. & Flaherty, K. T. BRAF targeted therapy changes the treatment paradigm in melanoma. *Nature Reviews Clinical Oncology* **8,** 426–433 (2011).
105. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27,** 66–75 (2009).
106. Rudd, M. L. *et al.* Mutational analysis of the tyrosine kinome in serous and clear cell endometrial cancer uncovers rare somatic mutations in TNK2 and DDR1. *BMC cancer* **14,** 884 (2014).
107. SAM/BAM Format Specification Working Group, The. The SAM/BAM Format Specification (v1.4-r985). Sourceforge.net. 9 Sep 2009. Web. 19 July 2013. <http://samtools.sourceforge.net/SAM1.pdf>
108. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **39,** D38–D51 (2010).
109. Scharer, C. D. *et al.* Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer research* **69,** 709–717 (2009).
110. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
111. Shen, M. M. & Abate-Shen, C. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes & Development* **24,** 1967–2000 (2010).
112. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* **26,** 1135–1145 (2008).
113. Sherry, S.T. *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311 (2001).
114. Shi, L. *et al.* Whole-genome sequencing in an autism multiplex family. *Mol Autism* **4,** 8 (2013).
115. Smith, N. G. C. Deterministic Mutation Rate Variation in the Human Genome. *Genome Research* **12,** 1350–1356 (2002).

116. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genetics* **41,** 393–395 (2009).
117. Stefansson, O. A. *et al.* A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular Oncology* (2014). doi:10.1016/j.molonc.2014.10.012
118. Takai, D. *et al.* Large scale mapping of methylcytosines in CTCF-binding sites in the human H19 promoter and aberrant hypomethylation in human bladder cancer. *Human molecular genetics*, **10**, 2619-2626 (2001).
119. Tervasmäki, A., Winqvist, R., Jukkola-Vuorinen, A. & Pylkäs, K. Recurrent CYP2C19 deletion allele is associated with triple-negative breast cancer. *BMC Cancer* **14,** 902 (2014).
120. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* **39,** D214–D219 (2010).
121. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
122. Tian, D. *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455,** 105–108 (2008).
123. Torkamani, A. & Schork, N. J. Prediction of cancer driver mutations in protein kinases. *Cancer research* **68,** 1675 (2008).
124. Urruticoechea, A. *et al.* Recent advances in cancer therapy: an overview. *Current pharmaceutical design* **16,** 3–10 (2010).
125. Vandin, F., Upfal, E. & Raphael, B. Algorithms and Genome Sequencing: Identifying Driver Pathways in Cancer. *Computer* **45,** 39–46 (2012).
126. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Research* (2011). doi:10.1101/gr.120477.111
127. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput Biol* **6,** e1000641 (2010).
128. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nature Communications* **4,** (2013).
129. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry* **55,** 641–658 (2009).
130. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics* **46,** 573–582 (2014).
131. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009).
132. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40,** D930–D934 (2011).
133. Wei, G.-H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO journal* **29,** 2147–2160 (2010).

134. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics* **46,** 1160–1165 (2014).
135. Weischenfeldt, J. *et al.* Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. *Cancer Cell* **23,** 159–170 (2013).
136. Wong, W. C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27,** 2147–2148 (2011).
137. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576 (2014).
138. Yip, K.Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome biology*, **13**, R48 (2012).
139. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).
140. Young-Xu, Y. & Chan, K. A. Pooling overdispersed binomial data to estimate event rate. *BMC Medical Research Methodology* **8,** 58 (2008).
141. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387 (2014).
142. Zhang, Z. D. *et al.* Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Research* **17,** 787–797 (2007).

Supplementary Material for LARVA

1. Pseudogene UTR, TSS, and promoter sites removal

Pseudogenes are known to be hotspots of artifacts in numerous genomics analyses (Guo 2014). Part of the reason lies in the fact that read mapping in pseudogenes might be complicated due to their context similarity with their parent genes. In order to analyze the mutation events in the pseudogene regions, we extracted all the pseudogenes from the Gencode annotation (version 19) and calculated the average mutation counts from the pooled samples in gene and pseudogene regions, and also the upstream and downstream 2kb region of all pseudogenes. Possibly due to the shorter length of the pseudogenes, a larger variance of the mutation rate was observed in the pseudogenes compared to the genes, although two-sided Wilcoxon test shows no significant difference (P = 0.453). However, we observed a noticeable elevated mutation rate in the upstream and downstream regions of pseudogenes (Fig. S2), which can potentially affect the UTR, TSS, and promoter regions analysis. In order to exclude potential artifacts, such as variant calling artifacts, we excluded the pseudogenes from the Gencode gene list when analyzing these regions.

2. Details of model fittings

2.1. The constant mutation rate assumption and the resultant binomial distribution

The underlying assumption of the binomial model used in Alexandrov *et al.* (2013) is that the mutation rate within the given region is a constant. Suppose the target region has n based in length, and the homogeneous mutation rate is p. Then

the mutation count x inside this region falls into a binomial distribution with the probability mass function as:

$$\Pr(x=k)=\binom{n}{k}p^k(1-p)^{n-k} \tag{1.1}$$

Given the mutation count data, the maximum likelihood estimator of the mutation rate is:

$$\hat{p}=\frac{\sum_{i=1}^{k}x_i}{\sum_{i=1}^{k}n_i} \tag{1.2}$$

where k represents the total number of regions and i is the region index.

2.2. The beta-binomial distribution used in LARVA

Instead of the fixed mutation rate assumption, we provided more flexibility of the mutation rate by allowing it to follow a beta distribution:

$$\begin{aligned}\pi(p|\alpha,\beta)=Beta(\alpha,\beta)&=\frac{p^{\alpha-1}(1-p)^{\beta-1}}{Beta(\alpha,\beta)}\\&=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\end{aligned} \tag{1.3}$$

Suppose the mutation count is $x_i, i=1,2,\ldots,k$, and the sample size and binomial probability can be expressed as n_i and p_i. Instead of assuming the mutation counts in all bins is a constant, we can set up a two stage model:

$$\begin{aligned}x_i|\,p_i &\sim Binomial(n_i,p_i)\\p_i &\sim Beta(\alpha,\beta)\end{aligned} \tag{1.4}$$

Then the total number of mutations within a bin of length n follows the beta binomial distribution as in (1.5):

$$\Pr\{X = x_i\} = \binom{n_i}{x} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x_i)\Gamma(\alpha+n_i-x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_i)} \tag{1.5}$$

To estimate the parameters in the beta-binomial distribution, we used the scheme described in Young-Xu & Chan (2008) and Kleinman (1975). When the target bin length is fixed, resulting in $n_i = n, i = 1,2,\ldots,k$, the mean and variance of mutation counts can be written as:

$$\begin{aligned} E[X] &= n\frac{\alpha}{\alpha+\beta} = n\mu \\ \mathrm{var}[X] &= n\mu(1-\mu)\sigma, \\ \sigma &= \frac{1}{\alpha+\beta+1} \end{aligned} \tag{1.6}$$

For simplicity, we directly estimate μ and σ instead of α and β. Hence, the moment estimator can be immediately derived from equation (1.6).

When the target region length is variable, the estimation is a little bit more complicated. We define additional parameters for mathmetical convenience in (1.7):

$$\begin{aligned} \hat{p} &= \frac{\sum_{i=1}^{k} w_i \hat{p}_i}{w} \\ w_i &= \frac{n_i}{1+\sigma(n_i-1)} \\ w &= \sum_{i=1}^{k} w_i \\ S &= \sum_{i=1}^{k} w_i (\hat{p}_i - \hat{p})^2 \end{aligned} \tag{1.7}$$

From this, we can derive the moment estimator in (1.8):

$$\mu = \hat{p} = 1 - \hat{q}$$
$$\sigma = \frac{S - \hat{p}\hat{q}\left[\sum_{i=11}^{k} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]}{\hat{p}\hat{q}\left[\sum_{i=1}^{k} w_i\left(1 - \frac{w_i}{w}\right) - \sum_{i=1}^{k} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]} \quad (1.8)$$

However, from (1.8), w_i is also a function of σ, which is to be estimated, and there is no analytical solution to it. Hence, as suggested in Kleinman (1975), we initially assigned the w_i proportional to n_i to get a rough estimate of γ. Then w_i was updated with this estimate to obtain a more accurate estimation of σ.

3. Coding Region Mutation Burden Analysis

LARVA is not designed for coding region analysis due to the availability of synonymous sites, which serve as a natural and biologically meaningful background in these regions. For the sake of gaining additional insight from an exome performance calibration, we nevertheless evaluated LARVA's ability to identify statistically significant mutation burdens in genes. Exome variant data was obtained from The Cancer Genome Atlas (TCGA) Data Portal (Cancer Genome Atlas Research 2008). The complete set of exome variant calls includes 20 cancer types and 5032 samples in total. A detailed graph of the collected data is provided in Fig S8.

Gene annotation data was derived from the GENCODE v19 annotation files (Harrow 2012). All complete protein coding transcripts were extracted, and all the exons for each gene were merged, as demonstrated in Fig S9. This data spanned 19,822 genes, and a total of 252,356,877 nucleotides. We plotted the distribution of gene lengths in Fig S10. The total number of mutations falling into the merged gene regions is 3,547,350, and the average mutation rate is 0.0141 for the pooled samples.

As with the noncoding regions, huge mutation heterogeneity was observed in the coding regions (Fig S11). Then we calculated the replication timing for each merged exon and gave the final replication timing for each gene as its exon length weighted average value.

We removed the genes with length less than the bottom 5% of gene lengths for higher annotation confidence, and then compared the performance of LARVA and the binomial test. After *p*-value adjustment, LARVA found 15 genes that are potentially under higher mutation burden (Table S2). For each of these genes, we searched for literature supporting cancer association. We found 11 out of 15 genes are clearly documented with some cancer association. Note that we reported only one Pubmed ID per gene, even if there are many more supporting references. Our findings effectively demonstrate that LARVA is capable of finding meaningful results on protein coding regions. On the other hand, the *p*-values for the binomial test method were heavily inflated. After *p*-value adjustment, there are 3,927 out of 19,110 genes, roughly 20.55%, with adjusted p-value less than 0.1. It is very unlikely that all such genes are associated with cancer.

Q-Q plots of *p*-values given by LARVA and the binomial test are given in Fig. S12. It is shown that the *p*-value distribution from the binomial test severely violates the uniform distribution assumption, which is consistent with its bad fitting of the data. On the other hand, the *p*-values from the LARVA method (Fig. S12, left hand side) roughly follow a uniform distribution.

4. Importance of covariate correction

It is well known that local mutation rate is affected by various factors, such as replication timing and GC content (Lawrence 2013). Due to these confounding factors, the observed mutation count data distribution is actually a mixture of several different distributions, which further increases the overdispersion. We used some simulations to show this effect in Fig S14. We randomly simulated five binomial distributions of binomial (n=100,000), where the mutation rate p_i was uniformly sampled from [$1e^{-6}$, $5e^{-5}$] to mimic the mutation rate difference from various replication timing regions. The empirical distribution of the pooled data is given in the pink line in Fig S14. It is shown that even if we ignore patient-specific heterogeneity, the observed data demonstrates much larger variation than expected simply by mixing several different binomial distributions. It is necessary to remove such effects for more reasonable *p*-value calculation.

5. Factors that affect overdispersion in the mutation count data

5.1 Heterogeneity in mutation rates in different patients/cancer types

Suppose that even if mutation rate per sample is constant across the whole genome, it may vary between different patients/cancer types. Each patient-specific mutation rate can be considered a random sample from beta distribution, so after pooling all samples together, the mutation counts for each bin follows a beta-binomial distribution. The overdispersion under this condition depends on how different these patients are.

5.2 Length of the target region to be analyzed

Assume that y is the number of somatic variants in n bases, the point mutation rate is ε. Unlike the constant mutation assumption in binomial distribution, we assume that ε is a random variable with:

$$E(\varepsilon) = p$$
$$Var(\varepsilon) = \varphi p(1-p)$$

So the variance of y can be calculated as:

$$\begin{aligned} Var(y) &= E_{\varepsilon}\left(Var\left(y|\varepsilon\right)\right) + Var\left(E\left(y|\varepsilon\right)\right) \\ &= E_{\varepsilon}\left[n\varepsilon(1-\varepsilon)\right] + Var(n\varepsilon) \\ &= n\left[p - \varphi p(1-p) - p^2\right] + n^2\varphi p(1-p) \\ &= np(1-p)\left[1 + (n-1)\varphi\right] \end{aligned}$$

The variance in the mutation count data is scaled by a factor of $1+(n-1)\varphi$.

Biologically speaking, the difference of the point mutation rate within the analyzed noncoding region varies much more in longer noncoding elements (2.5kb promoters) compared to smaller regions (200bp TSS), resulting in very different overdispersion parameters in the estimation stage. Hence, we do not recommend evaluation of regions of non-comparable length in the same run.

6. Supplementary figures

Figure S1: Boxplot of mutations counts in 10k, 100k, and 1mb regions with or without overlapping blacklist regions. P-values were calculated from the two-sided Wilcoxon tests. It is clear that in the smaller 10kb and 100kb regions the mutation rate is much larger in the blacklist bins compared to the nonblacklist bins. Due to the mappability issues surrounding blacklist regions, many of these intersecting variants are likely to be spurious, and hence they were excluded from the remaining analyses.

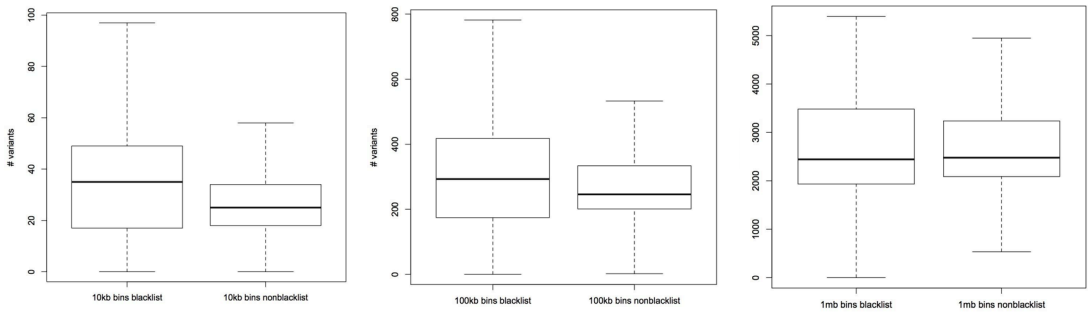


Figure S2: The distribution of mutation rates of all gene types (both protein coding genes and pseudogenes), protein coding genes only, pseudogenes only, and the 2kb regions upstream and downstream of pseudogene annotations. Pseudogenes exhibit a much larger mutation rate variance. If genes and pseudogenes were evaluated together, the extremely high mutation rate pseudogenes could be mistakenly classified as significantly mutated (false positives). Hence, the pseudogene annotations were removed from the LARVA gene analysis.

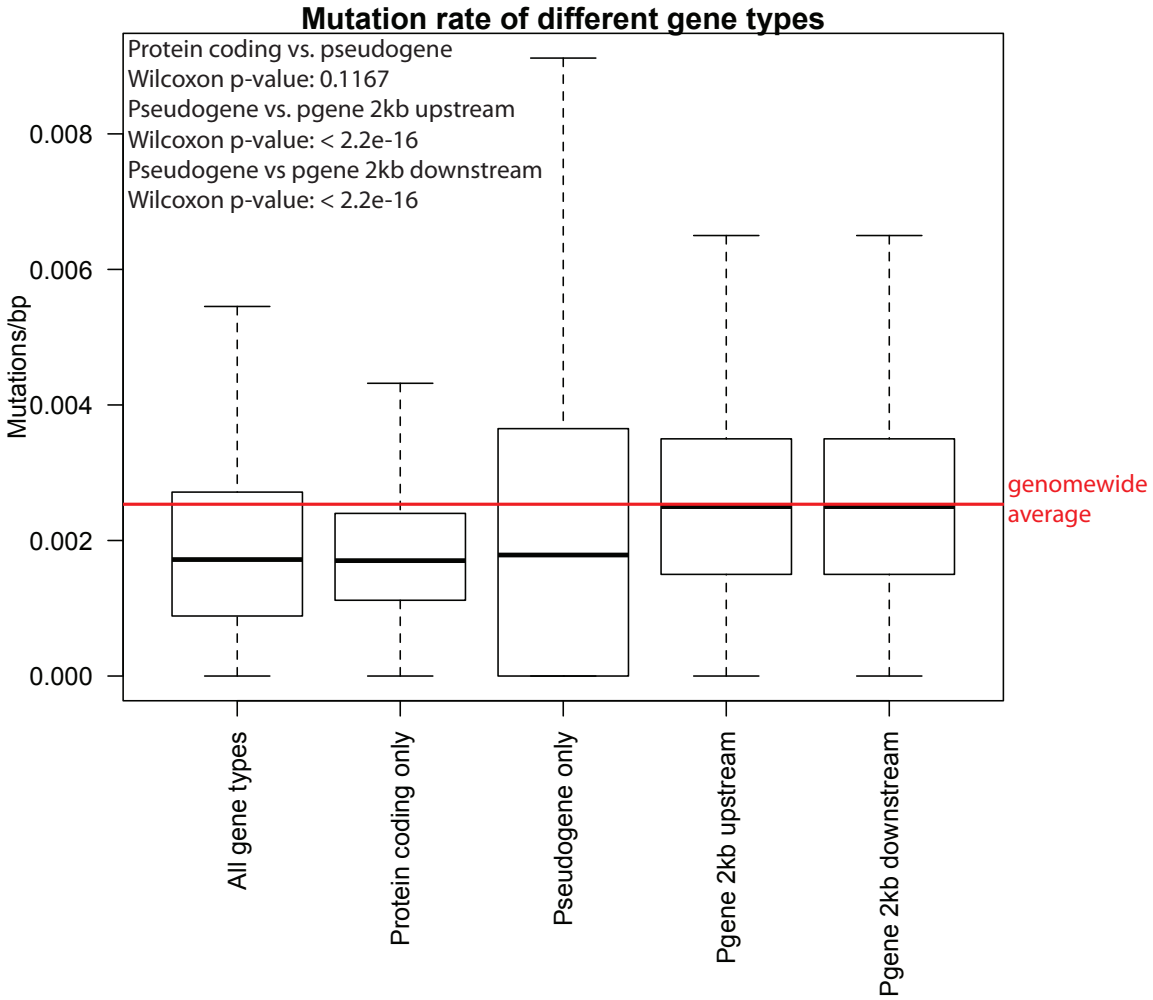


Figure S3: Fitting comparison between the beta-binomial and binomial distribution. (A) Density plot of the beta-binomial, binomial, and empirical distribution of read count data in 100kb bins; (B) C.D.F curve of the KS statistics of beta-binomial and binomial generated counts vs. random samplings in the observed counts; (C) Boxplots of the KS statistics.

Fig. S3

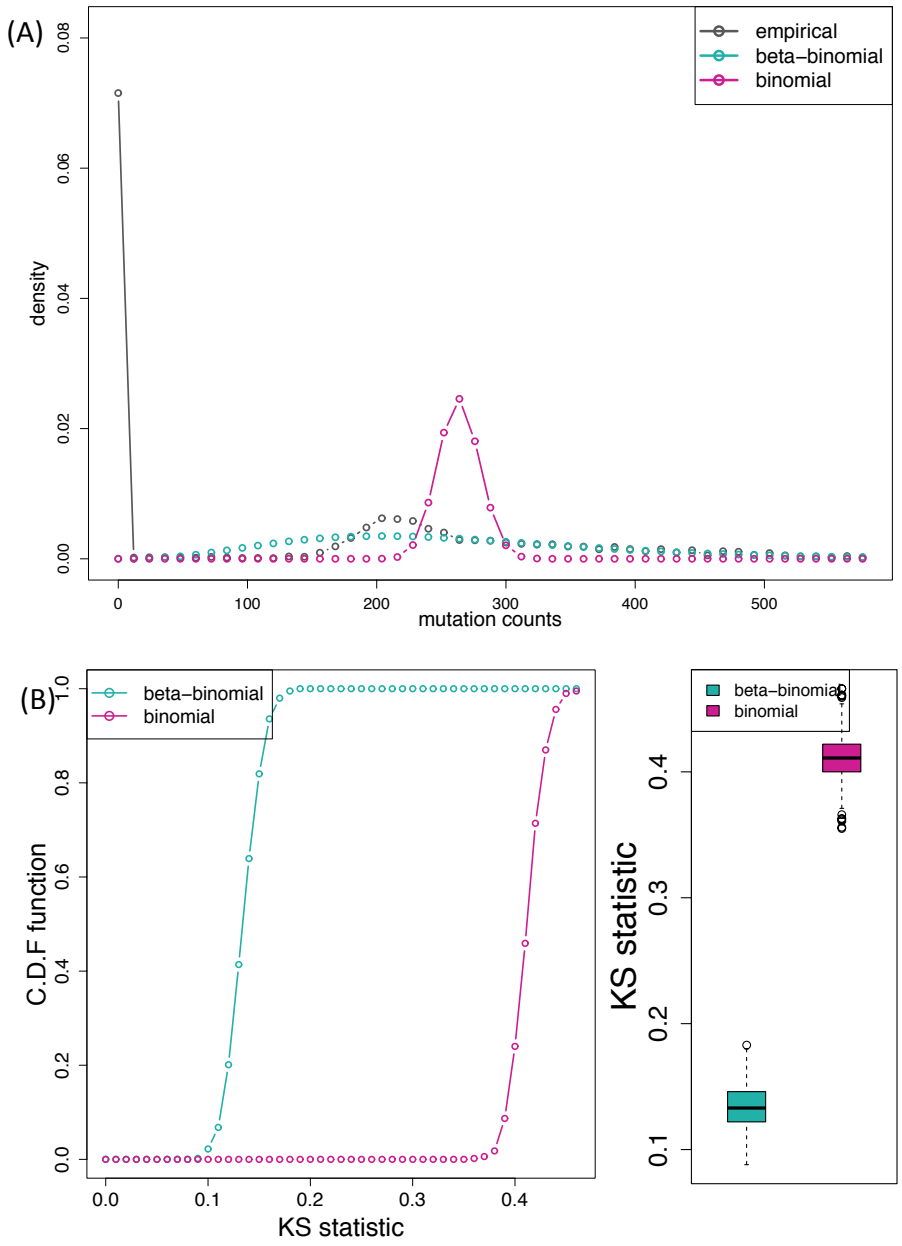


Figure S4: Fitting comparison between beta-binomial and binomial distribution. (A) Density plot of the beta-binomial, binomial, and empirical distribution of read count data in 1kb bins; (B) C.D.F curve of the KS statistics of beta-binomial and binomial generated counts vs. random samplings in the observed counts; (C) Boxplots of the KS statistics.

Fig. S4

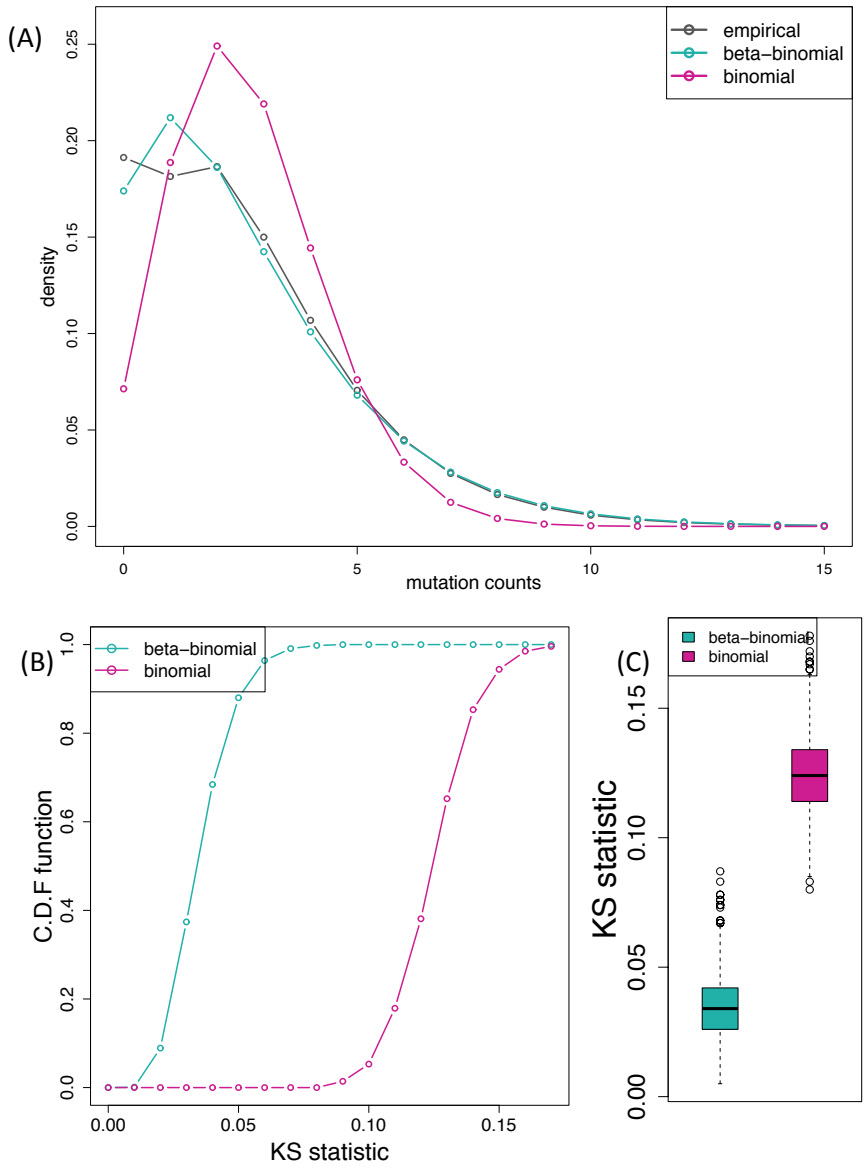


Figure S5: Half of the observed data is used for model fitting of both beta-binomial and binomial distribution, and the remaining half was used to calculate the KS statistics with generalizations from the fitted distributions. Boxplots of 100 repeats were given below.

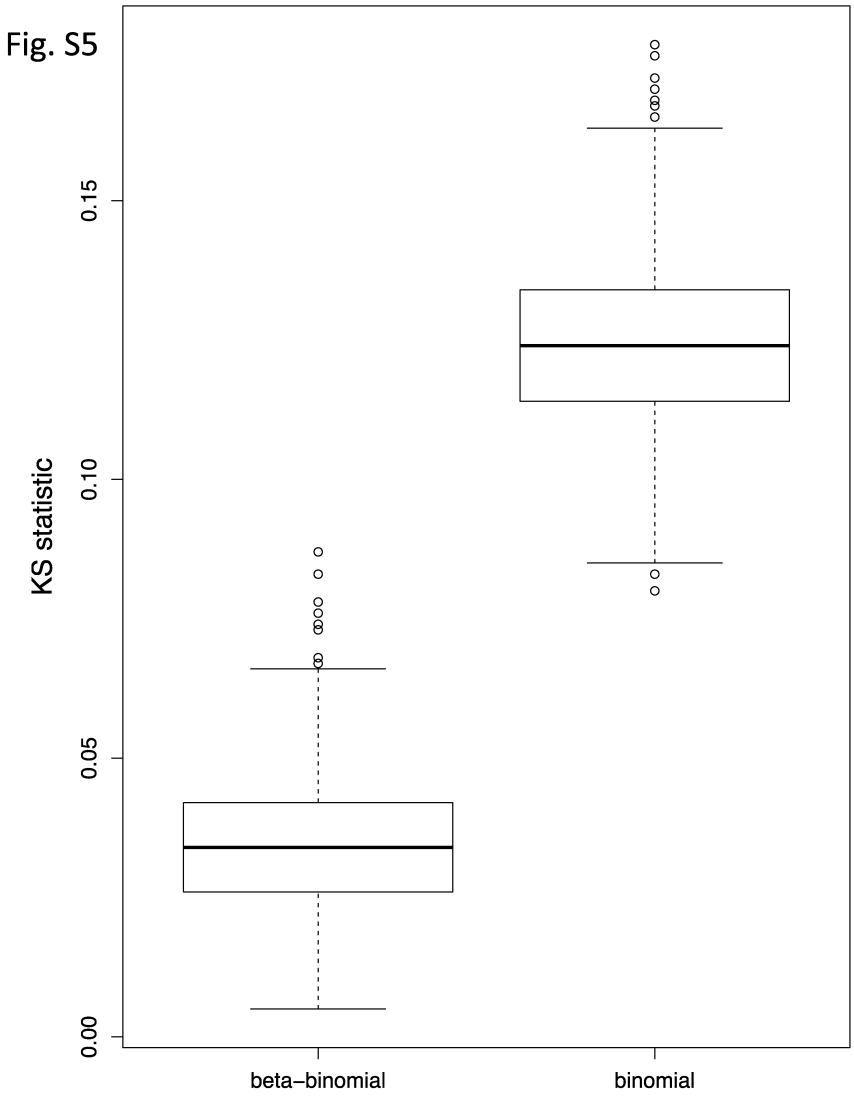


Figure S6: The smooth scatterplot of the mutations count in all tumor samples within 1kb bins vs. its averaged replication timing value. A linear regression was fitted and the R-squared values are up to 0.124.

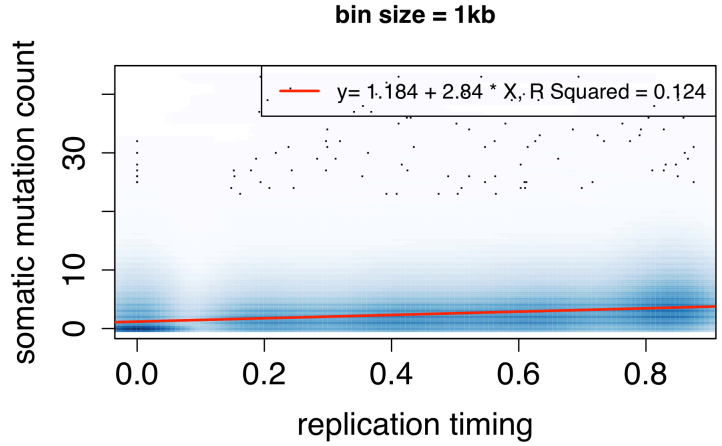


Figure S7: The fitted μ and σ were plotted for each the 10 used replication timing bins.

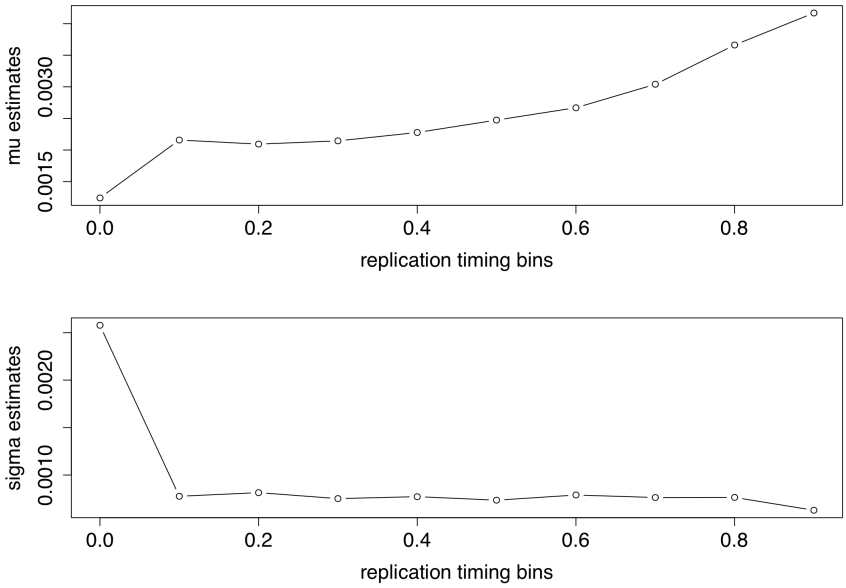


Figure S8: TCGA Whole Exome Sequencing samples by cancer types.

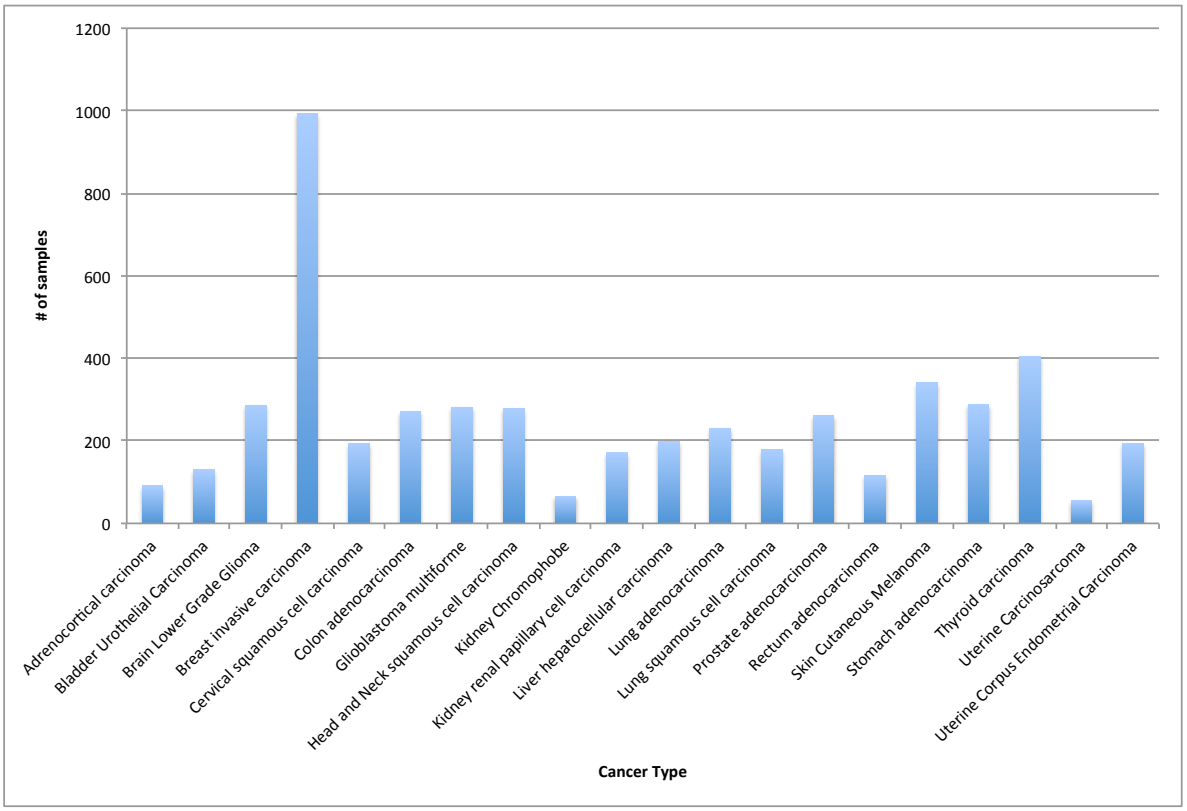


Figure S9: Details of gene regions definition. Note that only coding transcripts were used for the Whole Exome Sequencing data analysis.

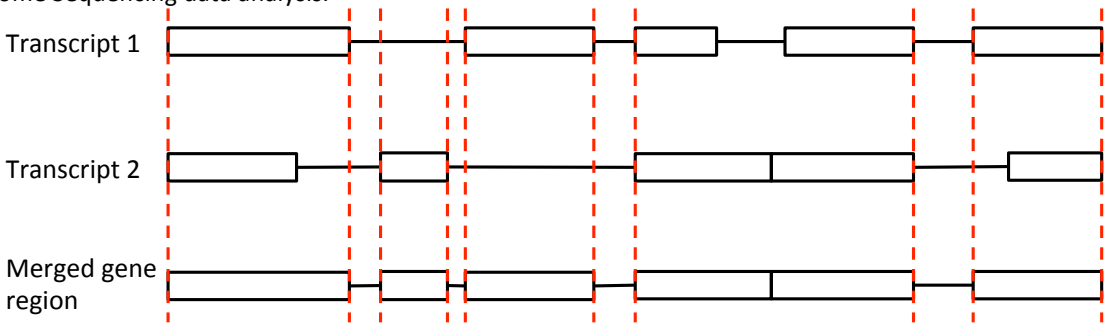


Figure S10: Distribution of gene length

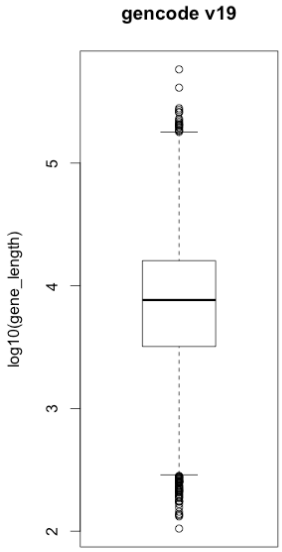


Figure S11: Distribution of the pooled mutation rates

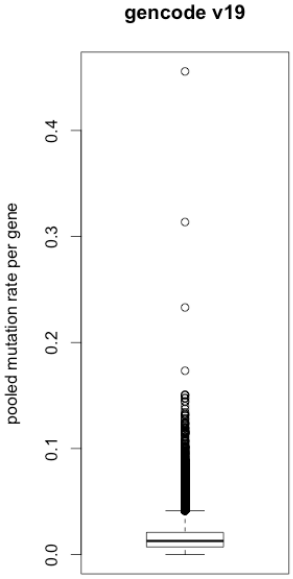


Figure S12: In the coding regions: A) The average mutation rate vs. replication timing; B) QQ plots of calculated p-values vs. uniform theoretical ones of the coding region analysis.

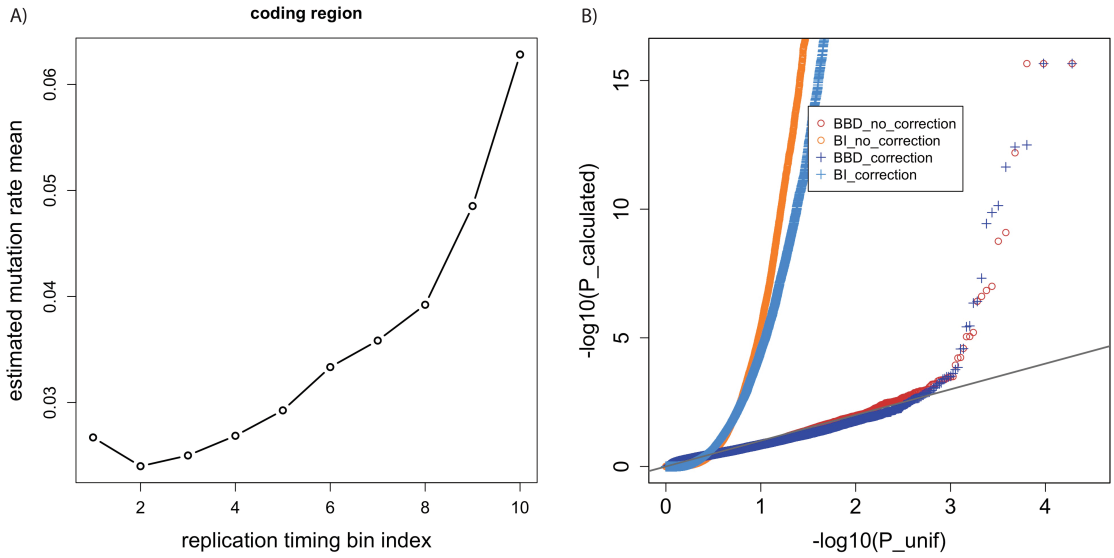


Figure S13: QQ plots of calculated p-values vs. uniform theoretical ones of the real and permuted datasets. Variants were derived from the pan-cancer dataset from Alexandrov *et al*. (2013), the prostate cancer variants from Baca *et al*. (2013), and the stomach cancer variants from Wang *et al*. (2014) "Real" refers to the original variant data, and "permutated" refers to the dataset created by randomizing the variant positions of the original dataset within a 25kb window centered on the original position.

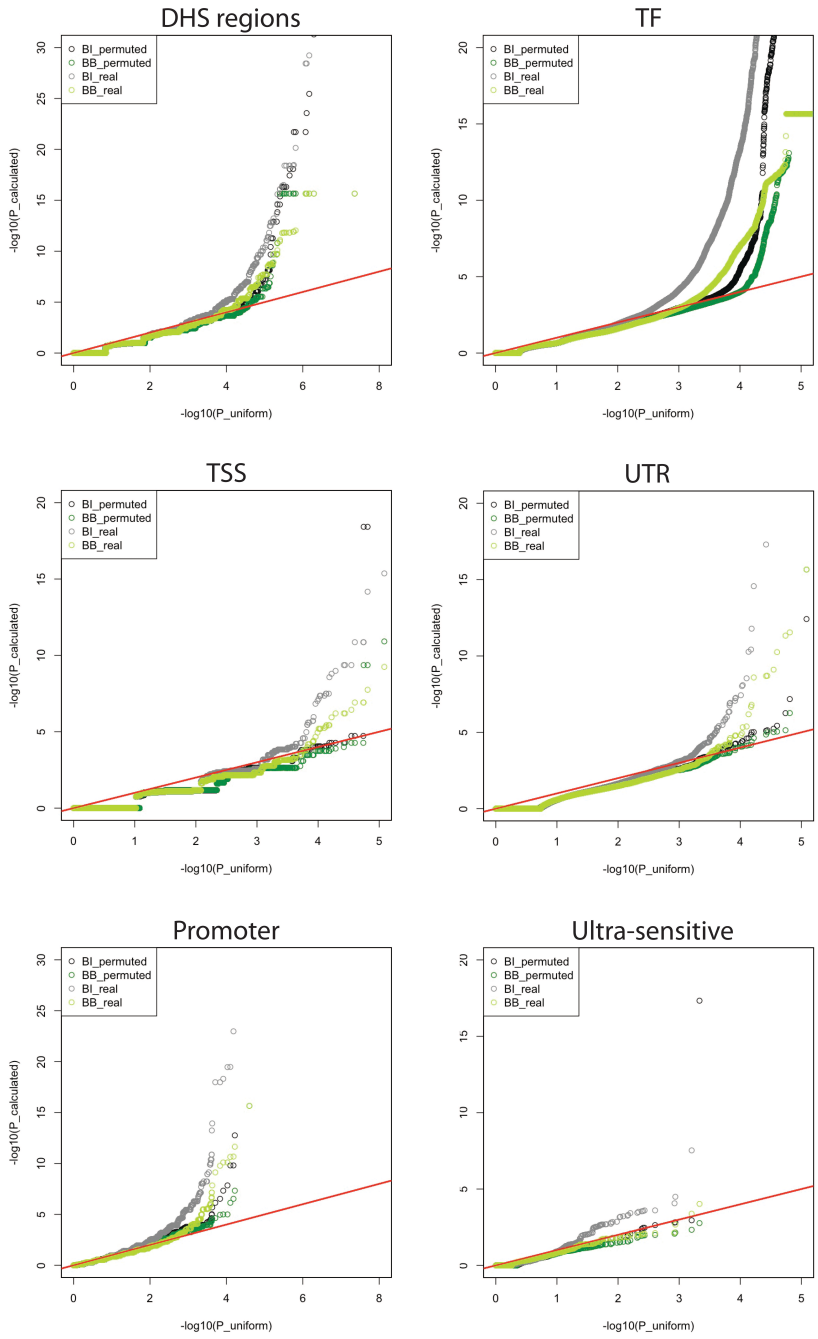
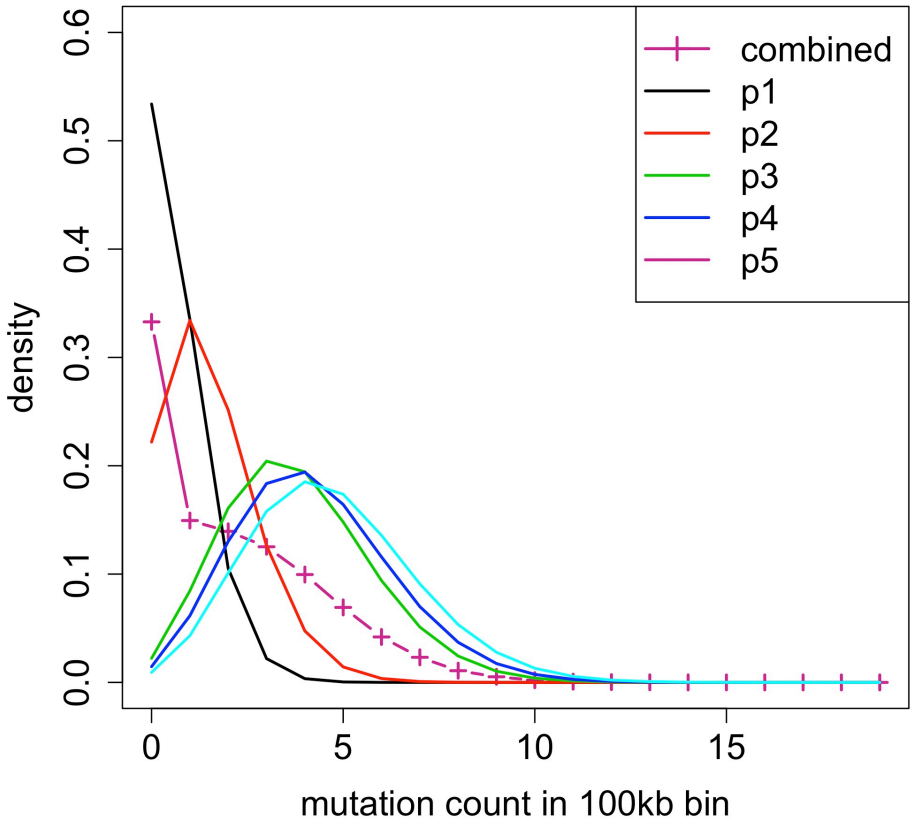


Figure S14: Importance of covariate correction. The pink line shows a mixture of five different binomial distributions. This mixture effect introduced extra variation of the mutation count data.



7. Supplementary tables

Table S1: Summary of the whole genome sequencing cancer data used for the LARVA study

Cancer Type	**# Samples**
Acute Lymphoblastic Leukemia	1
Acute Myeloid Leukemia	7
Breast Cancer	119
Chronic Lymphocytic Leukemia	28
Glial Tumor	26
Kidney Carcinoma	32
Liver Cancer	88
Lung Adenocarcinoma	24
Lymphoma B-cell	24
Medulloblastoma	100
Pancreatic Cancer	15
Pilocytic Astrocytoma	101
Prostate Cancer	95
Stomach Cancer	100
TOTAL	760

Table S2: Genes with significant mutation burden, according to LARVA's exome analysis. CGC stands for Cancer Gene Census.

Rank	Gene name	PubMed ID	Annotation
1	TP53	17401430	CGC listed, breast; colorectal; lung; sarcoma; adrenocortical; glioma; Spitzoid tumour; multiple other tumour types
2	KRAS	24755884	CGC listed, pancreatic; colorectal; lung; thyroid; AML; other tumour types
3	FRG1B	19586940 ,24084849	ALL (Acute lymphocytic leukemia)
4	NRAS	20619739	CGC listed, melanoma; MM; AML; thyroid
5	PTEN	18794879	CGC listed, glioma; prostate; endometrial
6	KRTAP4-11	NA	NA
7	PIK3CA	16449998	CGC listed, colorectal; gastric; glioblastoma; breast
8	IDH1	19935646	CGC listed, glioblastoma
9	BRAF	12068308	CGC listed, melanoma; colorectal; papillary thyroid; borderline ovarian; NSCLC; cholangiocarcinoma; pilocytic astrocytoma; Spitzoid tumour; pancreas acinar carcinoma; melanocytic nevus; prostate; gastric
10	B2M	18506145	Colorectal cancer
11	CDKN2A	14993899	CGC listed, melanoma; multiple other tumour types
12	TBP	12697807	TBP is up-regulated by oncogenic signaling pathways and may be a critical component in dysregulated signaling that occurs downstream of genetic lesions that cause tumors
13	KRTAP5-4	NA	NA
14	KRTAP5-5	NA	NA
15	KRTAP4-5	NA	NA

8. References

1. Young-Xu, Y. and Chan, K.A. (2008) Pooling overdispersed binomial data to estimate event rate. *BMC medical research methodology*, **8**, 58.
2. Kleinman, J.C. (1975) Proportions with extraneous variance: two dependent samples. *Biometrics*, **31**, 737-743.
3. Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.
4. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, **22**, 1760-1774.
5. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214-218.
6. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415-421.
7. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666-677.
8. Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S. *et al.* (2014) Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics*, **46**, 573-582.