B Significance

In this proposal we aim to prioritize rare, non-coding variants associated with cancer. This work represents a collaboration between a computational scientist (Mark Gerstein) and an experimental cancer genomicist (Mark Rubin). Gerstein and Rubin have worked together for most of the last decade, co-publishing many papers during that period.

B-1 Much recent progress in annotating the non-coding genome, making it ripe for variant annotation

Annotating non-coding regions is essential for investigating genome evolution [1], for understanding important biological functions (including gene regulation and RNA processing) [2], and for elucidating how SNPs and structural variations may influence disease [3]. Many projects related to annotating the noncoding genome have recently come to completion. The Encyclopedia of DNA Elements (ENCODE) Project recently provided a comprehensive catalogue covering much of the entire human genome [4]. In addition, the model organism ENCODE (modENCODE) Project presents an extensive genomic annotation of drosophila [5] and *C. elegans* [6] and a way to relate this to human. Furthermore, large-scale mRNA and miRNA sequencing have been applied to elucidate the functional landscape of regulatory variations in the human genome [7, 8, 9, 10]. Similar efforts have been directed toward annotating human epigenomic data to investigate underlying disease mechanisms [11]. Moreover, the important role of regulatory variants in various diseases have generated a great deal of interest in identifying and annotating the expression of Quantitative Loci linked to specific genes [12, 13].

B-2 Non-coding variants, most of which are regulatory, are significant to the study of diseases but less well studied than coding variants

Numerous studies have been conducted on the mutations to coding portions of the genome. However, comparatively less effort has been invested in the investigation of disease-related disruptions to noncoding portions of the genome. Nevertheless, a few initial studies indicate that variants in non-coding regions of genome significantly influence the associated phenotype [14] and are often implicated in various diseases[15, 16]. Much of the non-coding variation is contributed by regulatory variants, where cis- and trans-acting variation in the human genome can modulate gene expression [17] and this gene expression variation has been implicated in cancer and other diseases [18, 19, 20, 21, 22, 23]. Specific examples are expression quantitative trait loci (eQTLs) and variants associated with allele-specific behavior. It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription-factor (TF) binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states [24]. Genotype-transcript associations have been reported at large for multiple types of inherited variants [8, 9, 25, 26, 27], however experimental evidence of inherited variants allele-specific effect on enhancer/promoter activities and transcriptional influence (short and long range) are lacking.

B-3 Rare variants are significant to study of cancer & disease in general

There have been a large number of GWAS studies [28], which have primarily focused on the identification of common genetic variants. They have neglected the role of rare variants (particularly in noncoding regions) in various diseases [29]. However, growing evidence suggests that these rare genetic variants have strong effects and can act as a primary driver of many human diseases, including cancers [30]. Increased disease susceptibility is often attributed to the cumulative effect produced by multiple rare variants [31]. For instance, bioinformatic and biochemical analyses indicate that rare germline variants in the CHEK2 gene [32] and PALB2 gene increase the risk of breast cancer [33]. In addition, a rare variant (rs138212197) in the HBOX gene [34] and a rare SNP (rs188140481) in the telomeric region of the 8q24 locus were found to be associated with prostate cancer [35].

B-4 Rare variants in cancer patients in similar functional elements as somatic variants may be associated with disease risk

In cancer studies, particularly related to tumor sequencing, prior studies have primarily emphasized the identification of somatic over germline variants. For instance, the current TCGA call sets do not even contain "official" germline calls. However, rare germline and somatic variants have often been observed in the same genetic element across multiple individuals. The reciprocity between germline and somatic variants may increase the risk of cancer in such individuals, and we plan to identify these elements using data on large populations. Multiple experimental studies support this point of view. Germline and somatic mutations in the promoter region of the telomerase reverse transcriptase (TERT) gene have been observed in cutaneous melanoma [20]. Similarly, many somatic and germline mutations in the T53 gene and GALNT12 coding exons were implicated in Sonic-

Hedgehog medulloblastoma (SHH-MB) tumors [36] and colon cancers [37], respectively. The interplay between somatic and germline variants in hMSH6 and hMSH3 genes has been shown to be associated with gastrointestinal cancer [38]. A similar association was discovered between two germline SNPs and somatic mutations in the EGFR signaling pathway in colorectal cancer [39]. In recent years, there has been a growing interest in understanding the contribution of germline and somatic variants in tumor expression [18]. Similar studies have been proposed to investigate whether these associations augment the risk of triple-negative breast cancer and prostate cancer among African American populations.

C Innovation

Our method will combine various large-scale genomics data to interpret rare non-coding variants associated with increased cancer risk. Currently no computational pipeline exists with focused analysis for germline variants associated with increase cancer risk. Moreover, large-scale consortia, such as the 1000 Genomes and ENCODE, have produced data that can be used to interpret other genomic studies. However, these resources have not been fully exploited to understand the functional implications of variants associated with increased cancer risk. The integration of these data would be an important innovative component of our approach. The specific innovative components of our approach are listed below.

C-1 Identifying and interpreting rare non-coding variants associated with increased cancer risk using population-scale polymorphism data

The GWAS catalog contains many common variants associated with diseases. However, as discussed above, many rare variants increase cancer susceptibility. Currently, no standard methods exist to functionally interpret such variants, especially in non-coding regions. Thus, our approach will be the amongst the first for functional interpretation of these variants. The 1000 Genomes consortium has created a deep catalog of genetic variation across many populations. Our approach will use the allele frequencies of variants in ~2,500 individuals from 1000 Genomes data to understand which genomic regions are tolerant to common mutations without conferring disease risk. We will then use this knowledge to identify rare variants that may be associated with increased disease risk.

C-2 Using non-coding annotations to understand the likely biological role of non-coding variants

The ENCODE consortium has annotated non-coding regions of the genome. One of the major aims of these annotations is to help understand genetic variants that cause disease by misregulation of gene expression. Our approach will be innovative since it will be amongst the first methods that use ENCODE data to interpret variants that increase cancer susceptibility.

C-3 Using knowledge of somatic cancer-causing variants to identify germline variants associated with increase cancer risk

We will use knowledge of somatic variants that constitute cancer driver events to identify germline variants associated with increased cancer susceptibility. Thus, our approach will be innovative in analyzing somatic and germline variants in an integrative fashion.

C-4 Analyzing variants in ncRNAs

Most previous studies for functional interpretation of noncoding GWAS variants have primarily focused on regulatory regions associated with transcription factor binding sites or regions of open chromatin. Our approach will also analyze variants in ncRNAs and thus this will form another major innovative component of our approach.

C-5 Functionally validating rare variants

Rare variations in regulatory regions of genome can have a paramount influence on biological processes and might function as primer for recurrent somatic mutations in adjacent genomic regions or might contribute to long range changes in chromatin regulation. Using a comprehensive panel of cell lines and genome editing tools like the CRISPR-CAS system we introduce the rare variations in the cell lines and study effect on cellular physiology. This innovative approach will allow us to generate a catalogue of biological outcomes that can be attributed to a rare variation in a physiological setting.

D Approach

D-1 Approach Aim 1 - Convert the prototype FunSeq non-coding somatic variant pipeline to prioritize germline variants and elaborate it with new features

D-1-a Preliminary Results for Aim 1

D-1-a-i We have considerable experience annotating non-coding regulatory regions of the genome

Our proposed work is based on our experience in non-coding annotation. We have made a number of contributions in the analysis of the noncoding genome, as part of our extensive 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs [40], as well as new machine learning techniques [41]. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF's target genes based on ChIP-seq data [42]. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers [43], which we have partially validated [44]. We have also constructed regulatory networks for humans and model organisms based on the ENCODE [45] and modENCODE datasets [46], and completed many analyses on them[6, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]

D-1-a-ii We have considerable experience processing RNA-seq data and annotating ncRNAs

We also have extensive experience conducting integrated analyses of large sets of RNA-seq data, such as through the ENCODE, modENCODE, BrainSpan and exRNA consortia [4, 6, 61]. In particular, for general RNA-Seq analysis, we have developed RSEQtools, a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models [62]. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts [63]. Comparisons between RNA-Seq samples, and to other genome-wide data, will be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks [64]. We have also developed a ncRNA-finder [65]. Finally, we have developed statistical models relating gene expression levels to chromatin marks and TF binding [45, 66, 67, 68].

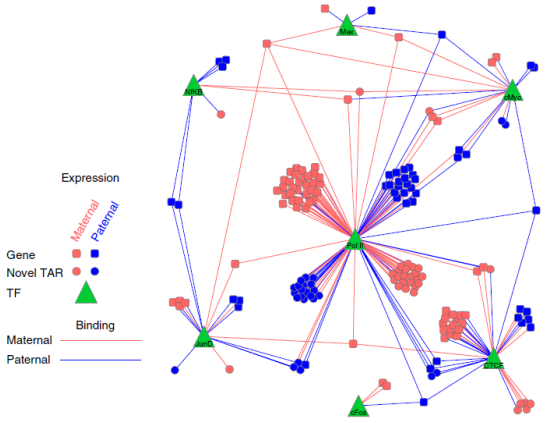

Fig 1: Network adapted from Rozowsky et al. (2011) \cite{21811232} depicting allele-specific regulation of the expression of genes and transcriptionally active regions (TARs) by binding of transcription factors (TFs). Edges represent regulation of TFs in allele-specific fashion to genes and TARs. Pink and blue denote maternal and paternal entities respectively. Circles represent TARs, squares genes and green triangles TFs used in publication. We would have many more TFs for pipeline than 7 shown here.

D-1-a-iii We have extensive experience in Allelic Analysis

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE) [69, 70]. We have previously developed a tool, AlleleSeq, [58] for the detection of candidate variants associated with ASB and ASE. Using AlleleSeq, we have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project. [45, 56, 61] Overall, we found that there is a substantial number of genomic elements associated with ASB and ASE [61] and that these allelic variants are under differential selection from non-allelic ones [45, 56]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression. [45] Furthermore, we have provided the AlleleSeq tool, lists of detected allelic variants, and the constructed personal diploid genome and transcriptome of NA12878 on \cite{alleleseq.gersteinlab.org}. Since then, we updated the AlleleSeq tool, and the resource has been used in the scientific community, as exemplified by the number of citations and publications using our data as references. [71, 72].

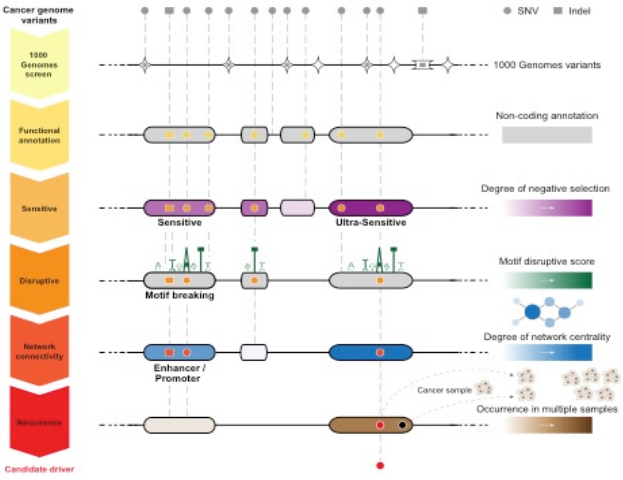

Fig 2: Detailed description of FunSeq workflow.

D-1-a-iv We have extensive experience in relating annotation to variation & based on this experience have developed the prototype FunSeq pipeline for Somatic Variants

We have extensively analyzed patterns of variation in non-coding regions along with their coding targets [44, 45, 73]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations [73]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region [4]. Further studies by our group showed relations between selection and protein

network structure, e.g. hubs vs periphery [55, 57].In a recent study [56], we have integrated and extended these methods to develop a prototype prioritization pipeline called FunSeq. FunSeq identifies sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation. It also prioritizes variants based on network connectivity and their disruptiveness (e.g. finding motif breakers) and identifies deleterious variants in many non-coding functional elements, including transcription-factor (TF) binding sites, regions of active chromatin corresponding to enhancer elements and regions of open chromatin corresponding to DNase I hypersensitivity sites. By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq allows for identification of candidate non-coding driver mutations [56]. In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples.

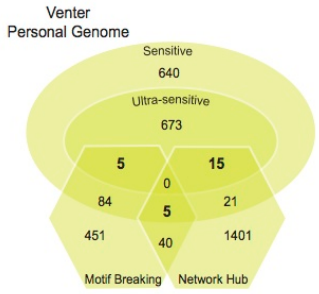

Fig 3: Application of variants filtering scheme to Venter's genome. Number of SNVs in various categories shown.

D-1-b Research Plan for Aim 1

We plan to convert the current FunSeq prototype from its focus on somatic variants to allow the identification of rare variants associated with high functional impact. We will do some simple improvements (i.e. incorporating GERP scores and ultra-conserved regions for identifying conserved regions between species) and some major changes outlined below.

D-1-b-i Identifying gain-of-function mutations for TF binding sites in addition to loss-of-motif events

Loss-of-function variants are more likely to cause deleterious impact [56, 73, 74]. When variants occur in TF binding motifs, the change in position-weight matrix (PWM) can be calculated. Variants decreasing the PWM scores could potentially alter the binding strength of transcription factors, or even cause loss-of-motif events. Many studies have shown that gain of new binding sites caused by somatic mutations can constitute driver events [19, 20, 75, 76]. However, an automated tool to detect such events in whole genomes is not available. Such events in germline genomes might also be associated with increased disease risk. We will create a gain-of-motif scheme to scan and statistically evaluate [77] all possible motifs created by variants compared to the human reference genome. Gain-of-motif events are identified as those that give a sequence score with mutated allele in the PWM significantly higher than the background. Note that in these analyses, determining the ancestral allele of the variant is essential to resolving between loss-of-function or gain-of-function since the functional impact of the variant reflects the historical event when the polymorphism was first introduced in the human population.

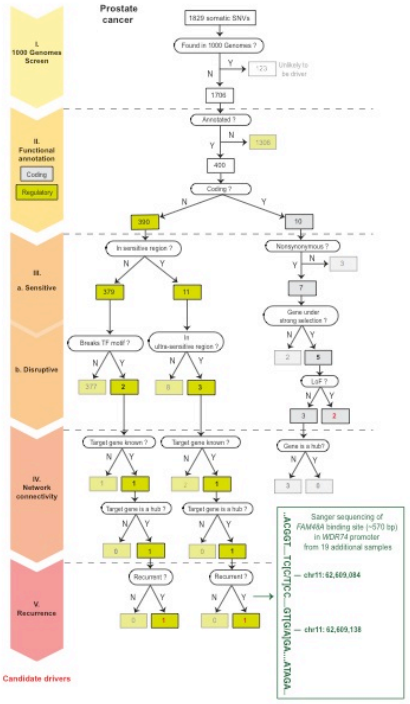

Fig 4: Filtering of somatic variants from a prostate cancer sample leading to identification of candidate drivers

D-1-b-ii Identifying likely target genes of distal regulatory elements & then assessing impact of variants on network connectivity

To interpret likely functional consequences of non-coding variants, we will define associations comprehensively between many non-coding regulatory elements and target protein-coding genes. We will consider the enhancer marks H3K4me1 and H3K27ac as two types of activity signals, and DNA methylation as an inactivity signal. We will collect all bisulfite sequencing, ChIP-seq and RNA-seq data from the Roadmap Epigenomics project [78]. Then we will identify significant associations between regulatory elements and candidate target genes through computing the correlations of active signals and anti-correlations of inactive signals with gene expression levels across different tissue types.

We will use the regulatory element - target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory the network, metabolic pathways, etc. We know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious [55, 57]. For each non-coding variant, we will calculate scaled network centrality (compute the percentile after ordering centralities of all genes in a particular network) of the associated gene in various networks. If the associated gene participates in multiple networks, we will use the maximum network centrality as the disruptive measure of the variant. In addition to hubs,

we will also prioritize based on bottlenecks and positions at the top of hierarchies [45]. We will make the scheme flexible so it can integrate user-supplied networks.

Moreover, the interpretation of the functional impact of variants can be enhanced if the function of its target protein-coding genes is known. We will incorporate prior knowledge of genes, such as known cancer-driver genes [79] and actionable genes ('druggable' genes) [80] into our annotation scheme. We will also make the scheme flexible so it can integrate gene expression studies in cancer cases vs controls to increase predictive power for identification of functional variants (e.g. using DESeq[81]).

D-1-b-iii Detailed variant prioritization for ncRNAs

The original FunSeq focused on TF binding sites. Here, we will expand FunSeq to better prioritize variants in ncRNAs, in a parallel fashion to what we have done for binding sites. We will first prioritize ncRNAs based on their within-human selection pressure and conservation across multiple species, identifying sensitive regions. For within-human selection, we will prioritize annotations showing higher nucleotide diversity and fraction of rare variants [73]. We will look at GERP scores [82] for inter-species conservation.

We will divide ncRNA annotations according to their subcategories, expression levels, and specificity of expression in cell lines. We will take into account subcategories including transfer RNAs, miRNAs, 5S ribosomal RNAs, small nucleolar RNAs, small nuclear RNAs, and long non-coding RNA. Expression levels of ncRNAs will be obtained from the ENCODE where RNAseq was performed on dozens of cell lines \cite{http://genome.crg.es/encode_RNA_dashboard/hg19}. We will prioritize ncRNAs that have higher expression levels and those that are ubiquitously expressed in many cell lines.

Furthermore, we will annotate genomic variants with secondary structures of ncRNAs. Our preliminary data have shown that more rigid structures, such as stem regions, are under stronger selection pressure, and that those variants that incur a larger free energy change of the structures tend to be rarer in the human populations. We will also quantify the effect a mutation stabilizes or destabilizes the RNA structure by computing the difference in folding free energy changes of the RNA before and after the introduction of the mutation. RNA secondary structures will be predicted using RNAshapes [83]. After we do this, we will be able to define RNA-disruptive variants analogously to how LOF variants are defined for coding regions or motif-breakers are defined for TF binding: we will define variants that disrupt secondary structures of ncRNAs as those that no longer form a complementary base-pairing or a wobble base-pairing when mutated. (Again the correct identification of the ancestral allele will be important here.)

Finally, we will explore the relationship of ncRNAs with network connectivity by associating ncRNAs with canonical genes through expression correlations, sequence complementarity, etc. For instance, miRNAs are known to regulate the expression level of its target genes. We will identify coding genes associated with miRNAs by correlating their expression levels based on RNAseq. In addition, we will also search for potential miRNA binding target by examining sequence complementarity in 3'UTR regions of coding genes to the seed regions, i.e. the first 2-7 bp of the mature miRNAs, using TargetScan [84]. We will then examine the selection pressure in ncRNAs that are associated with genes in network hubs vs. periphery.

D-1-b-iv Variant prioritization based on Allelic activity & eQTL association (AlleleDB module)

The evident regulatory roles of the allele-specific variants assert that they will be useful in identifying functional variants. However, currently, there is no prioritization scheme that integrates ASB and ASE regulatory variants. Previous analyses have been primarily variant-specific or focused mainly on a deeply sequenced individual, GM12878 [45, 56, 61]. Furthermore, an enrichment of rare variants among allelic variants [7] implies that a direct overlap of variants in a prioritization pipeline will not be applicable. (That is, we would not expect any of the allelic variants to directly overlap the rare variants prioritized by FunSeq.) Therefore, to enable the incorporation of allele-specific variants into the annotation pipeline, our strategy is to aggregate allelic variants into meaningful regions, or what we term 'allelic' genomic elements. We define 'allelicity' as the degree of how allele-specific a particular genomic element or category of elements is averaged over all the allelic variants in it. This is a continuous measure with a range of values as opposed to a binary variable of whether a variant is allele-specific. For example, an 'allelic' class of TF binding site might possess more allele-specific ASB variants, or a particular class of elements such as enhancers and promoters might be more allelic than another. In a similar vein, we also plan to extend this approach to integrate another category of regulatory variants: quantitative trait loci (QTL), such as Dnase I sensitivity QTLs (dsQTLs), splice QTLs (sQTLs) and expression QTLs (eQTLs). All the results will be housed in a central repository, which we called the AlleleDB. This will be used as part of the pipeline to prioritize variants, by up-weighting those input variants that are found in our list of allelic and eQTL elements.

D-2 Approach Aim 2 - Implement an efficient & easy to use FunSeq pipeline & run on all the germline variants in TCGA/ICGC

In this aim we will provide an efficient implementation of FunSeq, including a weighting system to bring together all its features, call all the rare germline variants in sequenced tumor genomes, and then run FunSeq on them to develop a prioritized variant and element list. Overall, using FunSeq prioritization plus screening out the common variants will allow us to identify the rare variant on a haplotype block with the greatest impact. We note that unlike GWA studies, which look for association signal, our method prioritizes variants based on functional information. Thus, the variants identified by our pipeline are most likely the causal variants. Furthermore, we will analyze the element-wise recurrence of these rare variants with somatic variants.

D-2-a Preliminary results in developing efficient tools & calling variants on a large-scale

We have significant experience in developing high-throughput tools for bioinformatics research. Our tools take the form of web services, distributed open source programs, annotation databases and distributed virtual machines. Many of the latter are hosted on Amazon Web Services Elastic Compute Cloud (AWS-EC2). In particular, for the analysis of high-throughput genomic experiments we have developed pipelines for analysing, RNA expression [85, 85, 86], alternative splicing [63], fusion transcripts [87], and copy-number variation [88]. We have developed pipelines for the analysis of regulatory networks [45, 47, 89, 89] and protein–protein interaction networks [50, 90, 91, 92, 93, 94].

We have much experience in large-scale germline variant calling through being active members of the 1000 Genomes Consortium, especially the Analysis Group and Structural Variant (SV) subgroups where majority of the variant calling tools are developed [95, 96, 97, 98]. We will use the Broad's Genome Analysis Toolkit (GATK) [99] for variant calling, which we have already extensively used previously [56]. *For rare variants, we will define them as variants not in 1000 Genomes (phase 1 or pilot) -- the "outersect" with 1000G -- as we did previously in the ENCODE production rollout [4, 56]*. Also, we will call some SVs, which are important contributors to human polymorphism, have high functional impact and are associated with disease [100, 101, 102]. We have developed a number of SV calling algorithms, including BreakSeq by comparing raw reads with a breakpoints library (junction mapping) [103], CNVnator by measuring read depths [104], AGE by refined local alignment [105], PEMer for paired ends [106], array-based approaches [107] and a sequencing-based bayesian model [108].

D-2-b Research Plan for Aim 2

D-2-b-i Do SNP & a limited amount of SV calling for all WGS Germline Variants in TCGA + ICGC

We currently have access to a combined >500 whole genome sequences from whole-genome sequencing of tumor-normal pairs (WGS) done by the Sanger Institute, The Cancer Genome Atlas (TCGA) [109], and various prostate cancer sequencing projects. (Most of this is available through dbGaP [110], to which we have obtained protected access subject to annual renewal.) We anticipate access to another ~2000 WGS genomes from International Cancer Genome Consortium (ICGC) [111] and TCGA. To call variants uniformly, we will run GATK with standard parameters on the TCGA+ICGC WGS results then filter the results. Additionally, we will run our CNVnator, BreakSeq and PEMMer software [103, 104, 106] on this data to identify copy number variants. We will filter against 1000 Genomes Phase 1 to define a pool of rare variants. We estimate within a single WGS the total germline SNPs will be ~3 million total variants and ~100,000 rare variants (in addition to ~10,000 somatic variants) and that each genome will take ~1 hr to process on our parallel cluster.

D-2-b-ii Analysis of recurrent germline & somatic variants (LARVA module)

We will develop a model to study the recurrence of both germline variants and somatic mutations across multiple cancer patients. We will aim to see if there are prioritized germline variants that affect the same element as somatic ones, in different individuals. On a simple level, recurrence would be a variant at exactly the same position in two individuals. However, this is exceedingly unlikely for rare or somatic variants [97]. Thus, we will consider mutational burden spread over elements, which include transcribed features, regulatory features, and groups of genes related through a common pathway or protein interaction subnetwork.

Our mutation recurrence discovery procedure has three stages. Given a cancer patient cohort, we will first identify recurrences in the somatic variants. We will then do the same for the rare, germline variants. The third step involves looking for connections between the two sets: elements that contain recurrent somatic variants and rare germline variants imply that the germline variant may be functionally connected with respect to cancer. The absence of common variants from these elements would serve as further evidence for a functional connection to cancer. We have developed a computational framework for identifying these types of recurrent variation, named Large-scale Analysis of Recurrent Variants and Annotations (LARVA). Given a set of cancer

patient whole genome variant calls, and a set of genome annotations, LARVA will pick out the recurrent variants, recurrently mutated annotations, and recurrently mutated subsets of annotations.

LARVA also has a module for computing the statistical significance of its results by simulating the creation of WGS variant calls with randomized variant positions. These random datasets, which otherwise contain the same number of samples and variants, are used to determine the null distribution of variants across the annotation set for comparison with the actual variant data. LARVA determines the positions of variants for its random variant datasets using a null mutation model designed to reflect the factors affecting the neutral mutation rates of different genome regions, and represents an extension of an exome null mutation model developed for MutSig [112]. These factors include the genome-wide DNA replication timings, since later replicating regions are more error-prone due to the depletion of free nucleotides [113]. Histone marks for H3K4me1 and H3K4me3 are used because they are anti-correlated with SNV density [114]. Also included is the whole genome RNA-seq data from the ENCODE project [4], representing the connection between expression and transcription-coupled repair [115]. Finally, the SNV density data from the 1000 Genomes Project [97] is used to reflect differences in genome regions' levels of natural population variation. The whole genome weight function is defined over discrete 100,000-bp-long regions of the genome, and is defined as follows for each region r:

weight(r) = log(CDF(r.replication_timing)) + log(1-CDF(r.H3K4me1)) + log(1-CDF(r.H3K4me3)) + log(1-CDF(r.expression)) + log(CDF(r.SNV_density))

Individual variant positions are selected by first choosing a region according to this weight function, then picking a position within that region with uniform probability.

D-2-b-iii We will implement FunSeq on a large scale & then run on all the variants to produce a shortlist of prioritized variants

D-2-b-iii-1 We will modularize FunSeq to handle updates to a complex data context & simultaneously carry out efficient production runs

We will develop a practical implementation of all of the new FunSeq modules proposed in aim 1 and then integrate them within FunSeq. Some of the modules may be useful as stand alone programs. For instance, for AlleleDB, the results will both be integrated into the pipeline and also housed in a standalone AlleleDB database. This can be navigated via a user-friendly interface for data mining and the casual user. It will also generate flat files for their queries and can be subsequently downloaded by the users for further analyses.

Our implementation will allow us to modularize FunSeq into two components: (#1) building a complex-to-regenerate data context and (#2) an efficient and high-throughput production run. To build the data context (#1), we will integrate large-scale publicly available data resources, such as polymorphisms from 1000 Genomes project [98], conservation data from [116, 117], functional genomics data from ENCODE [4] and REMC [78].

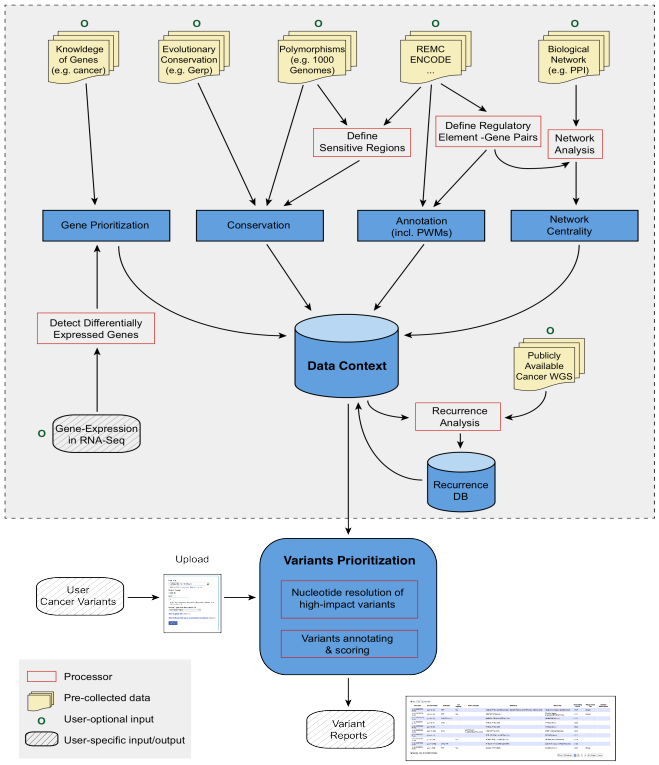

Fig 5: Description of extended FunSeq workflow.

We anticipate this step will be very time-consuming, as we will process large scale genomic data into smaller summary files (e.g. associations between distal regulatory elements and likely target genes). The production run (#2) will prioritize variants from WGS based on the data context. The variant prioritization step needs to be quite efficient, so we can tackle >1000 genomes in fairly short time. The overall modularization offers a flexible framework to for users to incorporate the ever-increasing amounts of genomic data to both rebuild the underlying data context and prioritize case-specific variants. We plan to make FunSeq an easy to use tool. It will be implemented as a downloadable tool, a web server, and a cloud instance.

D-2-b-iii-2 We will develop a unified weighted scoring scheme for combining all FunSeq modules to consistently prioritize variants

An integral part of the modular nature of FunSeq will be a way to combine the results of all of the modules into a

single variant score and consistent ranking. For this we will develop a weighted scoring system. Different features may contribute differently to the deleterious impact of variants. We will use the mutation patterns observed in the 1000 Genomes polymorphisms to assign weight values to features. Features that are frequently observed in polymorphisms will be weighted less, whereas those that are rarely observed will be weighted more. In general, features can be classified into two classes: discrete (e.g. "in a particular functional annotation or not") and continuous (e.g. the PWM change in 'motif-breaking'). We plan to use different strategies for the two classes. For each discrete feature d, we will calculate the probability $p_d$that it overlaps a natural polymorphism. Then we will compute 1-Shannon entropy as its weighted value w_d. This measure ranges from 0 to 1 and is monotonically decreasing when p_d is between 0 and 0.5.

$$w_d = 1 + p_d * \log_2 p_d + (1 - p_d) * \log_2(1 - p_d) \quad (1)$$

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in polymorphisms. Thus one weight cannot suffice. For a continuous feature c, which is associated with a score v_c(e.g. PWM change), we will calculate feature weights for each v_c. In particular, we will discretize at each value and compute $w_c^{v^c}$ using (2). Now, when we come to evaluate the continuous feature c for a particular variant, we calculate its weighted value using the actual v_c corresponding to the variant.

$$w_c^{v^c} = 1 + p_c^{\geq v^c} * \log_2 p_c^{\geq v^c} + (1 - p_c^{\geq v^c}) * \log_2(1 - p_c^{\geq v^c}) \quad (2)$$

Finally, for each cancer variant, we will score it by summing up the weighted values of all its features. We will also consider the dependency structure of features when calculating the scores.

D-2-b-iii-3 We will run FunSeq & Larva on all the variants & prioritize them

We will run FunSeq on the rare variants resulting from our variant calling on all the TCGA/ICGA whole-genome sequences. We expect ~100K per genome and for those variants to recur at the exact same position only rarely; thus, we will generate a prioritized list ~100M variants. We expect each rare variant to be on its own rare haplotype block; moreover, since we are explicitly screening out common variants, we expect only infrequently that there will be other variants on the same block. If we have multiple rare variants, we would expect FunSeq to differentially prioritize them, making it relatively straightforward to identify the "functional" SNP in each block. This situtation contrasts with what one observes in prioritizing relatively "common" GWAS SNPs, where finding the "functional" SNP in a block is a major challenge. From this pool of ~100M prioritized variants, we will select those in the top quartile that also recur in same element as a somatic variant in another individual, based on LARVA analysis. We will further prioritize variants with germline recurrence in the same element. Overall, this analysis will yield a list of the top 200 variants and elements associated with them. (Note this might not be exactly 200 elements, since it is possible that some of the same variants recur in the same element.) We will select 100 unique elements from this list and move them onto validation as described below.

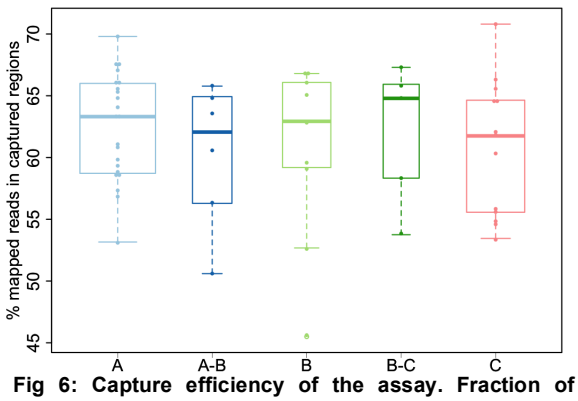

Fig 6: Capture efficiency of the assay. Fraction of reads in captured regions is reported, according to quality of the samples, ranging from highest quality (A) to lowest quality (C).

D-3 Approach Aim 3 - Validate the Prioritized Variants

D-3-a Preliminary results related to validation

D-3-a-i Capture-Seq identifies rare physiologically relevant mutations

We have applied hybrid capture technology to sequence specific regions with high coverage. Specifically, we have developed a novel targeted next-generation sequencing (NGS) assay, suitable for FFPE and frozen material. The developed protocol is as follows. DNA is extracted from 3x1.5mm FFPE cores, using the Promega Maxwell 16 system. DNA quality is determined using Agilent FFPE derived DNA quality assessment kit in a subset of cases. TruSeq library preparation is obtained using 1µg input DNA. Custom capture is performed using the NimbleGen SeqCap EZ library kit. Paired-end sequencing (2x75bp) is then performed using Illumina HiSeq 2500. Samples are multiplexed (5-7 samples per lane) to ensure a nominal coverage of ~25-40M paired-end reads per samples. Raw sequences are aligned to the human genome reference sequence (GRC37/hg19). This initial mapping is then refined following a series of computational steps to

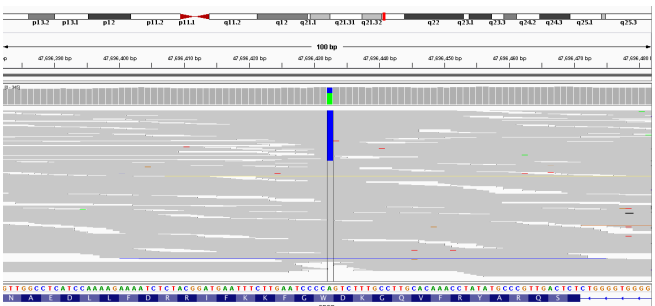

Fig 7: Integrative Genome Viewer (IGV) snapshot of SPOP mutation. Plot shows reads (grey) in the region of the SPOP gene where mutation was detected using hybrid capture. Colored bases identify variations with respect to reference genome. At mutation site total coverage is 280x with 187 (67%) and 93 (33%) reads supporting reference (A) and mutated allele (C), respec-

remove potential artifacts and increase the quality of the alignment. We then identify the somatic single nucleotide variants by comparing the tumor against its matching normal. In our study, we analyzed 31 cases of localized prostate cancer. We generated a total of ~1.340B paired-end reads (average per sample ~24.4M; range: 0.97M – 74M).The average coverage per sample is ~177x (range: 3x – 510x). The average capture efficiency is 61.4% (range 45.5% - 70.8%; see Figure 6). These results suggest that it is feasible to obtain good coverage with archival material with this assay. We were able to identify the known mutations in these samples, including TP53 and SPOP (see Figure 7), and to nominate some new ones. In this study, we were successful validating genomic alterations in samples up to 10 years old.

D-3-a-ii Low-frequency functionally active intronic & intergenic inherited variants predisposing to cancer

Emerging insights into the genetics of constitutional disease etiology demonstrate that germline polymorphisms are associated with a variety of diseases including Alzheimer's, Parkinson's, mental retardation, autism, schizophrenia [118]and cancer [119, 120]. Relevant to this proposal our group recently performed a large scale profiling study for 2,000 individuals from the Tyrol Early Prostate Cancer Detection Program [121, 122]cohort. This cohort is part of a population-based prostate cancer-screening program started in 1993 and intended to evaluate the utility of intensive PSA screening in reducing prostate cancer specific death. By genotyping DNA extracted from peripheral blood samples, we annotated the cohort on more than 5,000 CNVs and 900,000 SNPs and then queried inherited low frequency deletions variants [123] for their impact in driving prostate cancer [124] and the more aggressive form of the disease [125]. We reported on coding and non-coding functionally active risk variants. Among the top hits of the case-control study, an intronic variant in the *Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase C (MGAT4C)* demonstrated transcript abundance association with genotype states both in prostate and in lymphoblastoid cells, significant increase in cell and migration upon overexpression in benign and cancer prostate cell lines, and significant decrease in proliferation upon knock down of *MGAT4C* expression with siRNA. In addition, we suggested that intergenic PCA risk variants affect gene regulation through modified transcription factor binding activity of the Activator Protein 1 (AP-1) [24, 26]. Altogether, we demonstrated that inherited variants may directly or indirectly modulate the transcriptome machinery of known oncogenic pathways in prostate cancer facilitating carcinogenesis.

	H3K4me1	H3K4me1 + H3K3me3	H3K27ac	H3K9ac	DNase	FAIRE	UNION
AR	373 (136)	183 (55)	283 (98)	258 (83)	127 (39)	52 (16)	418 (148)
ER	386 (102)	221 (60)	317 (82)	339 (90)	232 (56)	127 (32)	431 (113)
AR+ER	17 (7)	9 (4)	14 (7)	17 (7)	6 (2)	3 (1)	22 (8)

Table 1: SNPs from human genome that intersect regulatory regions bound by AR and/or ERα.

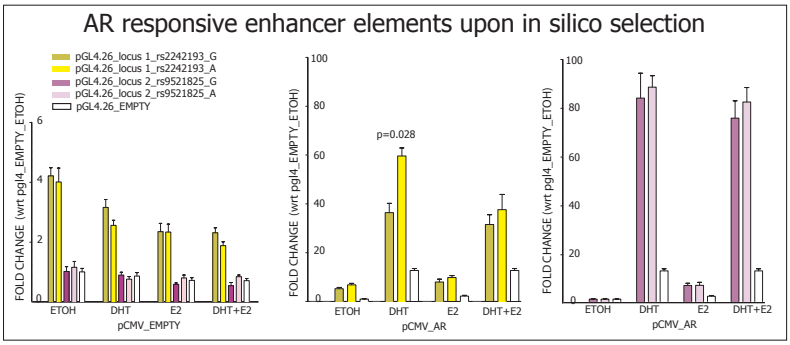

Fig 8: Validation AR-responsiveness and impact of SNPs within identified putative enhancer elements. Left panel: Cells cotransfected with pCMV_EMPTY vector and the different pGl4.26 constructs. Only pGl4.26_locus 1_rs12242193_G/A reaches as much as 4 fold change. Center panel: Cells cotransfected with pCMV_AR vector and pGl4.26_locus 1_rs12242193_G/A. Results show that the SNP has a role in transcription regulation when cells are supplemented with 100nM DHT (p=0.028, determined by Student's t-test). Right panel: Cells cotransfected with pCMV_AR vector and pGl4.26_locus 2_rs9521825_G/A. When cells are supplemented with 100nM DHT the construct reaches as much as 80-fold hinting at strong enhancer role. All experiments performed from three biological replicate, each one consisting of three technical replicates. Error bars indicates standard deviation of the mean (SD).

D-3-a-iii In vitro characterization of SNPs within enhancer elements bound by AR and/or ERα

The Tyrol Early Prostate Cancer Detection Program cohort is a well characterized cohort with centralized data collection that ensures proper patients' follow-up annotations and availability of well-preserved tissues and blood samples. The cohort currently includes more than 3,000 men. As part of our Trento-Innsbruck-Cornell collaboration, we further studied the genetics of prostate cancer individuals coupling serum levels and genomics data. Specifically, we studied the impact of genetic variants relevant to the metabolism of Dihydrotestosterone [126](DHT), the most potent form of androgen, and investigated the incidence of common genomic rearrangements with respect to PSA levels and age at diagnosis [127].

It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription factor binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states [24]. Genotype-transcript associations have been reported at large for multiple types of inherited variants [8, 9, 25, 26, 128], however experimental evidence of inherited variants allele-specific effect on enhancer activity are lacking. In order to study the potential role of inherited genetic variants within regulatory elements in the context of hormone dependent human, we have performed an unbiased computational search for AR/ERα bound enhancers elements containing SNPs followed by *in vitro* characterization of selected variants. **Table 1** shows counts of SNPs from the dbsnp137 set within AR [129] and/or ERα (Chakravarty D, *submitted*) binding sites that intersect peak ENCODE data [4] generated from 20 cell-lines and ChIP-seq experiments for H3K4m1, H3K4me1+H3K4me3, H3K9ac, H3K27ac, Dnase-seq and FAIRE-seq. For each marker the consensus was generated as the merge of all the regions that are present in at least 2 cell lines and comply with a set of filters. **Figure 8** shows examples of AR-responsiveness and SNPs impact on putative enhancer elements in MCF7 cells (Garritano S, Demichelis F, *unpublished*).

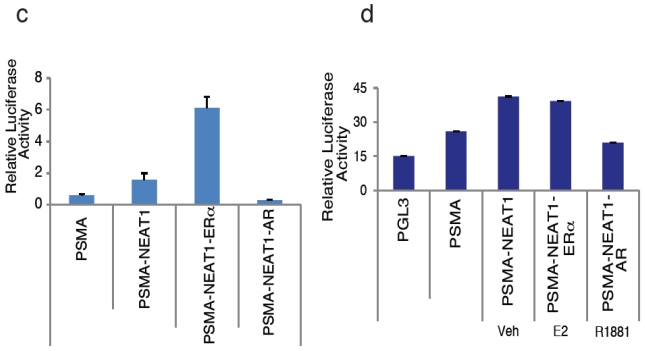

Fig 9:Promoter luciferase assay were performed using promoter reporter vectors encompassing ERα binding regions upstream of NEAT1. Luciferase assays confirmed ERa is recruited and drives transcriptional output from NEAT1 promoter

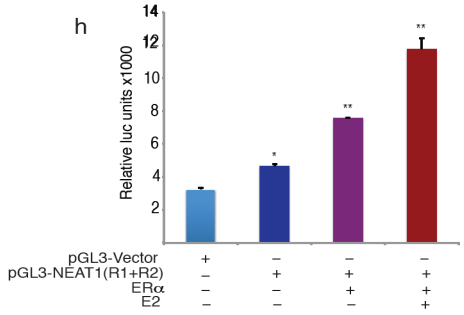

Fig 10: Promoter luciferase assay confirm NEAT1-ERα axis is involved in regulation of PSMA, a key prostate cancer gene.

D-3-a-iv Reporter luciferase assays confirm validity of in silico TF binding sites

Using and *in silico* approach we determined genome wide distribution of ERα in prostate cancer. Intriguingly, we observed a robust recruitment to non-coding genome and identified several intergenic sites that correlated with high ERα occupancy. Analysis of recruitment vs transcript profiles confirmed that ERα recruitment was associated with productive transcription of long noncoding RNA. Recruitment of ERα upstream of NEAT1 lncRNA was addressed in greater details. Reporter assays using promoter luciferase constructs encompassing upstream regulatory regions of NEAT1 and corresponding to two ERα binding sites are described in Fig 7.Interestingly, we discovered that NEAT1 is associated with chromatin and regulates transcription of key prostate cancer genes. Recruitment of NEAT1 was evaluated by ChIP assay and influence on key target genes like PSMA was validated using ChIP and reporter assays (Fig 8). Functional validation of NEAT1 functions revealed a predominant tumorigenic role as overexpression of NEAT1 was sufficient to augment proliferation, invasion and migratory behavior of prostate cancer cells (Fig 9).

D-3-b Research Plan Related to Validation

D-3-b-i Overview of validation strategy

Identification of rare variants and understanding the influence thereof on repertoire of biological responses will afford us a unique opportunity to understand causal role of these variations on other somatic mutations associated with diseased states including but not limited to cancer. The functional role of prioritized targets will be evaluated using a panel of cell lines that will serve as invitro-model to simulate effects *in vivo.* Once tested in cell line model we expect to extend these studies further to animal. We will use prostate cancer as a model for the validation but we expect that the results will be generalizable to a number of cancers.

First, we would perform an initial screen to determine whether any of the variants are associated with cancer in a different cohort of individuals or are associated with differential gene expression and RNA-seq. We will use both the Tyrol cohort (described above) and the Early Detection Research Network (EDRN) \cite{http://edrn.nci.nih.gov/} prostate cancer cohort with thousands

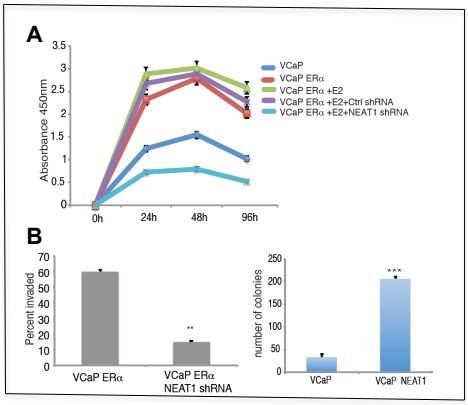

Fig 11: NEAT1 is a driver of oncogenic cascade (A) Cell proliferation assays were performed in VCaP and VCaP ERα expressing cells transfected with control or NEAT1 siRNA and estrogen treatment (10nM). (B) (left) Quantitative bar chart for depicting the relative cell counts obtained at the completion of the invasion assay performed in VCaP ERα control and NEAT1 shRNA expressing cells, (**) p < 0.01. (right) Soft agar assays were performed with VCaP control and NEAT1 expressing cells. Quantitative bar-plot analysis of stained colonies at 21

of prostate cancer individuals as well as normal controls. The prostate cancer cohort include men enrolled at three sites as part of the Prostate Cancer Clinical Validation Center that prospectively enroll individuals at risk for prostate cancer at Beth Israel Deaconess Medical Center (Harvard), at the University of Michigan (Michigan) and at Weill Cornell Medical College (Cornell). Cases are defined as men diagnosed with prostate cancer and controls are men who have undergone prostate needle biopsy without any detectable prostate cancer and no prior history of prostate cancer.

We will first take the highest prioritized variants then subject them to validation. Overall we plan to start the validation pipeline with the top ~100 elements identified from the computational FunSeq prioritization (as described above). We will perform Hybrid capture assay (described in preliminary data), on 400 cases (patients with cancer) from the above-mentioned cohorts. From the Capture-Seq experiments, we will identify the top 100 recurring variants and subsequently perform TaqMan assays on a further 4,000 cases to see if the precise variants recur in a larger cohort. From this group, we will select top third of the variants (~33), based on recurrence, that we will follow up for detailed functional screening, to be discussed below. This functional screening will be through various reporter assays (e.g. luciferase) looking for the effect on the target gene and also from using the CRISPR/Cas system. For controls, we will utilize deeply sequenced control cohorts (individuals with no cancer) that are already available, including deeply sequenced trios from the 1000 Genomes Project , 500 individuals with Complete Genomics sequencing also from 1000 Genomes and non-cancerous individual from the UK10K project \cite{http://www.uk10k.org/}.

D-3-b-ii Targeted sequencing & Genotyping

We will conduct the hybrid capture technology (as described in preliminary results) to sequence the top-ranking ~100 elements in 400 samples with high coverage. Custom capture will be performed using the NimbleGen SeqCap EZ library kit followed by paired-end sequencing (2x75bp) using Illumina HiSeq 2500.

Then we will utilize robust Taqman genotyping assays for screening ~100 nominated variants associated with the top-ranked elements in a cohort of 4000 individuals (Tyrol + EDRN, as described above). Superior allelic discrimination is achieved in these assays as they utilize TaqMan minor groove-binding (MGB) probes. This technique generates a low signal to noise ratio and affords a greater flexibility. The Taqman probes are functionally tested to first ensure assay amplification and optimization for amplification conditions.

Methods: Genomic DNA will be extracted from the blood cellular-EDTA samples in a high-throughput fashion using the QIAamp 96 DNA Blood Kit (Qiagen). All DNAs are evaluated by NanoDrop spectrophotometer (NanoDrop, Thermo Scientific) and gel electrophoresis (2% agarose). For TaqMan Real-Time Quantitative PCR, each DNA sample will be diluted to 10 ng/ml with nuclease-free water.

D-3-b-iii Evaluation of functional consequence of variants

Based on the Taqman results, we will pick the top third of the variants (~33) for functional follow up.

D-3-b-iii-1 Functional consequences: RNA-seq

First, we will use RNA-seq. We have RNA-seq data for many members of the cohort. To fill out the dataset, further RNA sequencing will be done on the cases where we see recurrent variants (on up to ~160 individuals). The RNA-seq will be done according to the protocols in [130]. This analysis will inform us if a SNP (in promoter or enhancer regions) has any effect on transcription of target gene. This analysis will provide a comprehensive list of SNPs that might correlate with loss or gain of expression. Recurrent rare SNPs will be further validated by PCR assays using primers that can amplify the genomic region encompassing the SNP. PCR will be followed by direct sequencing of amplicon using an ABI 3730 DNA Sequence Analyzer on a subset of tumor-normal pairs to verify the individual promoter/enhancer mutations for further confirmation.

D-3-b-iii-2 Functional consequences: Reporter Assays

Reporter assays that employ either LUC or next generation reporter vectors can provide direct insight to functional relevance of SNPs on target gene. GeneCopoeia offers Gaussia-luciferase (GLuc), eGFP,or mCherry based lentiviral or non-viral promoter reporter clones. In addition, we can also purchase Gluc vectors that are efficient tools to study transcription regulation. Minimal essential promoter region for each WT target gene will be subcloned from germline DNA using TOPO cloning kit (Invitrogen). If patient sample that harbors the mutation is available, we will amplify the corresponding mutant promoter sequence from the genomic DNA of the patient. PCR products will be cloned upstream to pGL-3-LUC promoter reporter plasmid or upstream to Gluc vectors. For each WT DNA Target gene-promoter plasmid a corresponding MT DNA Target gene-promoter plasmid will be generated using site directed mutagenesis utilizing QuikChange Lightning (Agilent). In this way we will have 33 WT promoter plasmids and 33 MT promoter plasmids in both PGL-3 LUC and Gluc background. We will utilize a panel of adherent cell lines. Cells will be seeded in 6 well plates and transfected with promoter reporter WT and mutant plasmid constructs. 48 hrs after transfection promoter activity will be meas-

ured following manufacturer's instructions. Assay values will be normalized using internal renilla luciferase as control.

Our expectation is that i*n vitro* promoter LUC assays will inform us if a particular mutation had any effect on transcription.

D-3-b-iii-3 Functional consequences: CRISPR/CAS system

We will utilize the newly discovered CRISPR/CAS system \cite{http://www.crispr-cas.org/} to generate endogenous mutations in target genes in a panel of cell lines. This unique system will provide us an opportunity to directly modulate endogenous genes and minimize artifacts due to the transfection based reporter assays. Using CRISPR/CAS mediated genome-engineering method [131] we will directly generate mutations within promoter/enhancers of target genes. Theoretically we generate 33 individual SNPs in each cell line and will study functional relevance of these changes compared to WT. In case of rare mutations, which occur within both promoter and enhancer regions of the same gene, we will develop cell lines having these combinatorial mutations. Mutations within regulatory regions like promoter and enhancer regions might contribute to one or more biological effects as described in the schematic. In addition to loss or gain of cognate coding transcript, it is quite conceivable that the SNPs might alter expression of non-coding transcript. To capture the complete influence of rare nominated SNPs at genomic and transcriptomic level we will perform RNA seq. The schematic shown represents representative iterations of plausible genomic changes that will be captured in this validation.

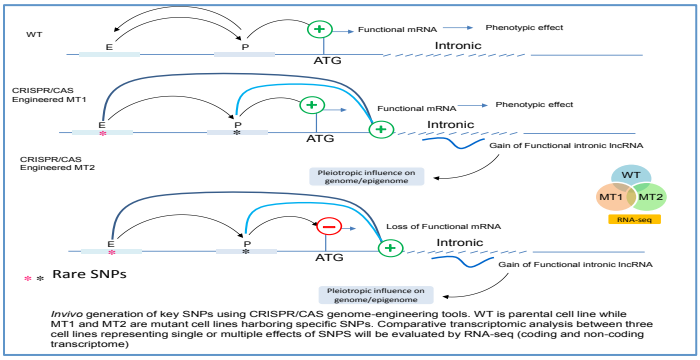

Fig 12: Overview of CRISPR/CAS system

Our expectation is that mutant and WT cell lines generated using CRISPR/CAS system will be monitored for a) phenoytypic changes by confocal microscopy and actin staining to determine effects of mutation on cytoskeletal reorganization b) Influence on proliferation by MTT and CellTiter-Glo® Luminescent Cell Viability Assay (Promega) c) Influence on invasive and migratory potential using, matrigel coated invasion and boyden chambers in 24 well format d) senescence by Bgal staining e) apoptosis by tunnel assay

D-3-b-iii-4 Functional consequences: Effect of the mutation on TF binding

In vitro EMSAs will confirm specific binding to WT or mutant sequence by a particular transcription factor. EMSA (electrophoretic mobility shift assay) is a common technique employed to study protein-DNA interaction. We will use the WT and the MT sequences to determine binding to a transcription factor predicted to be present at the site of mutation.

Chromatin immuno-precipitation (ChIP) assays for TFs overlapping the variant will be conducted to determine if the variant can distort TF binding. This would help validate the variants that are predicted to be motif breakers. Alternatively for the SNVs predicted to create a new motifs, ChIP experiments will help validate binding.

Timeline

Year I	Aim 1: Development of extended Funseq pipeline for annotating noncoding variants Aim 2: Optimization & beginning of variant calling Aim 3: development of validation assays
Year II	Aim 2: Germline variants called from ICGC/TCGA data Aim 2: Prioritization of most variants for validation experiments Aim 3: Begin functional validation experiments
Year III	Aim 2: Finishing prioritization of variants Aim 3: Functional annotation of prioritized variants Aim 2: Interpreting validation results in light of prioritization

1. Ponting, C. P. & Lunter, G. (2006) Signatures of adaptive evolution within human non-coding sequence. *Hum Mol Genet* 15 Spec No 2, R170-5.

2. Ghildiyal, M. & Zamore, P. D. (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10, 94-108.
3. Kleinjan, D. A. & van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* 76, 8-32.
4. ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.
5. modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., & Kellis, M. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330, 1787-97.
6. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, 3rd, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rätsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., & Waterston, R. H. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330, 1775-87.
7. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X., Dermitzakis, E. T., & Geuvadis Consortium (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506-11.
8. Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., & Dermitzakis, E. T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773-7.

9. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768-72.
10.Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., & Koller, D. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24, 14-24.
11. Won, K.-J., Zhang, X., Wang, T., Ding, B., Raha, D., Snyder, M., Ren, B., & Wang, W. (2013) Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res* 41, 4423-32.
12. Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 24, 408-15.
13. Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6, e1000888.
14. Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., & Gottesman, M. M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525-8.
15. Ward, L. D. & Kellis, M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 30, 1095-106.
16. De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., de Jong, P., Cheng, J.-F., Rubin, E. M., Wood, W. G., Bowden, D., & Higgs, D. R. (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312, 1215-7.
17. Cheung, V. G. & Spielman, R. S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 10, 595-604.
18. Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S., & Freedman, M. L. (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 633-41.
19. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science* 339, 957-9.
20. Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D., & Kumar, R. (2013) TERT promoter mutations in familial and sporadic melanoma. *Science* 339, 959-61.
21. Tournamille, C., Colin, Y., Cartron, J. P., & Le Van Kim, C. (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* 10, 224-8.
22. McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Altshuler, D., Daly, M. J., & Xavier, R. J. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* 40, 1107-12.
23. Verlaan, D. J., Ge, B., Grundberg, E., Hoberman, R., Lam, K. C. L., Koka, V., Dias, J., Gurd, S., Martin, N. W., Mallmin, H., Nilsson, O., Harmsen, E., Dewar, K., Kwan, T., & Pastinen, T. (2009) Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res* 19, 118-27.
24. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., & Snyder, M. (2010) Variation in transcription factor binding among humans. *Science* 328, 232-5.
25. Banerjee, S., Oldridge, D., Poptsova, M., Hussain, W. M., Chakravarty, D., & Demichelis, F. (2011) A computational framework discovers new copy number variants with functional importance. *PLoS One* 6, e17539.
26. Schlattl, A., Anders, S., Waszak, S. M., Huber, W., & Korbel, J. O. (2011) Relating CNVs to transcrip-

tome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res* 21, 2004-13.
27. Harris, W. H. (1992) Will stress shielding limit the longevity of cemented femoral components of total hip replacement?. *Clin Orthop Relat Res* , 120-3.
28. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106, 9362-7.
29. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012) Five years of GWAS discovery. *Am J Hum Genet* 90, 7-24.
30. Pritchard, J. K. (2001) Are rare variants responsible for susceptibility to complex diseases?. *Am J Hum Genet* 69, 124-37.
31. Bodmer, W. & Tomlinson, I. (2010) Rare genetic variants and the risk of cancer. *Curr Opin Genet Dev* 20, 262-7.
32. Sodha, N., Mantoni, T. S., Tavtigian, S. V., Eeles, R., & Garrett, M. D. (2006) Rare germ line CHEK2 variants identified in breast cancer families encode proteins that show impaired activation. *Cancer Res* 66, 8966-70.
33. Tischkowitz, M., Capanu, M., Sabbaghian, N., Li, L., Liang, X., Vallée, M. P., Tavtigian, S. V., Concannon, P., Foulkes, W. D., Bernstein, L., WECARE Study Collaborative Group, Bernstein, J. L., & Begg, C. B. (2012) Rare germline mutations in PALB2 and breast cancer risk: a population-based study. *Hum Mutat* 33, 674-80.
34. Ewing, C. M., Ray, A. M., Lange, E. M., Zuhlke, K. A., Robbins, C. M., Tembe, W. D., Wiley, K. E., Isaacs, S. D., Johng, D., Wang, Y., Bizon, C., Yan, G., Gielzak, M., Partin, A. W., Shanmugam, V., Izatt, T., Sinari, S., Craig, D. W., Zheng, S. L., Walsh, P. C., Montie, J. E., Xu, J., Carpten, J. D., Isaacs, W. B., & Cooney, K. A. (2012) Germline mutations in HOXB13 and prostate-cancer risk. *N Engl J Med* 366, 141-9.
35. Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Masson, G., Agnarsson, B. A., Benediktsdottir, K. R., Sigurdsson, A., Magnusson, O. T., Gudjonsson, S. A., Magnusdottir, D. N., Johannsdottir, H., Helgadottir, H. T., Stacey, S. N., Jonasdottir, A., Olafsdottir, S. B., Thorleifsson, G., Jonasson, J. G., Tryggvadottir, L., Navarrete, S., Fuertes, F., Helfand, B. T., Hu, Q., Csiki, I. E., Mates, I. N., Jinga, V., Aben, K. K. H., van Oort, I. M., Vermeulen, S. H., Donovan, J. L., Hamdy, F. C., Ng, C.-F., Chiu, P. K. F., Lau, K.-M., Ng, M. C. Y., Gulcher, J. R., Kong, A., Catalona, W. J., Mayordomo, J. I., Einarsson, G. V., Barkardottir, R. B., Jonsson, E., Mates, D., Neal, D. E., Kiemeney, L. A., Thorsteinsdottir, U., Rafnar, T., & Stefansson, K. (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* 44, 1326-9.
36. Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P. A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Bender, S., Pleier, S., Cin, H., Hawkins, C., Beck, C., von Deimling, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Blake, J., Benes, V., Kulozik, A. E., Witt, O., Martin, D., Zhang, C., Porat, R., Merino, D. M., Wasserman, J., Jabado, N., Fontebasso, A., Bullinger, L., Rücker, F. G., Döhner, K., Döhner, H., Koster, J., Molenaar, J. J., Versteeg, R., Kool, M., Tabori, U., Malkin, D., Korshunov, A., Taylor, M. D., Lichter, P., Pfister, S. M., & Korbel, J. O. (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* 148, 59-71.
37. Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., Lutterbaugh, J., Lawrence, E., Lewis, S., Willson, J. K. V., Lowe, J. B., Wiesner, G. L., Parmigiani, G., Barnholtz-Sloan, J., Dawson, D. W., Velculescu, V. E., Kinzler, K. W., Papadopoulos, N., Vogelstein, B., Willis, J., Gerken, T. A., & Markowitz, S. D. (2009) Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A* 106, 12921-5.
38. Ohmiya, N., Matsumoto, S., Yamamoto, H., Baranovskaya, S., Malkhosyan, S. R., & Perucho, M. (2001) Germline and somatic mutations in hMSH6 and hMSH3 in gastrointestinal cancers of the microsatellite mutator phenotype. *Gene* 272, 301-13.
39. Wormald, S., Milla, L., & O Connor, L. (2013) Association of candidate single nucleotide polymorphisms with somatic mutation of the epidermal growth factor receptor pathway. *BMC Med Genomics* 6, 43.

40. Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., & Gerstein, M. B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27, 66-75.
41. Yip, K. Y. & Gerstein, M. (2009) Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics* 25, 243-50.
42. Cheng, C., Min, R., & Gerstein, M. (2011) TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* 27, 3221-7.
43. Yip, K. Y., Alexander, R. P., Yan, K.-K., & Gerstein, M. (2010) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One* 5, e8121.
44. Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., & Gerstein, M. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13, R48.
45. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91-100.
46. Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R. F., Chen, J., Hwang, L., Cheng, C., Auburn, R. P., Davis, M. B., Domanus, M., Shah, P. K., Morrison, C. A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N. A., Grundstad, A. J., Hanley, D., MacAlpine, D. M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R. L., Ren, B., Posakony, J. W., Kellis, M., & White, K. P. (2011) A cis-regulatory map of the Drosophila genome. *Nature* 471, 527-31.
47. Cheng, C., Yan, K.-K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., Snyder, M., & Gerstein, M. (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* 7, e1002190.
48. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R. P., & Gerstein, M. (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* 107, 9186-91.
49. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., & Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet* 20, 227-31.
50. Yu, H., Zhu, X., Greenbaum, D., Karro, J., & Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* 32, 328-37.
51. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 3, e59.
52. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308-12.
53. Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., Bork, P., & Gerstein, M. B. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* 106, 1374-9.
54. Yu, H., Paccanaro, A., Trifonov, V., & Gerstein, M. (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* 22, 823-9.
55. Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* 104, 20274-9.
56. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüs, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacAr-

thur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., 1000 Genomes Project Consortium, Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C., & Gerstein, M. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342, 1235587.
57. Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 9, e1002886.
58. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M., & Gerstein, M. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7, 522.
59. Lörcher, U., Peters, J., & Kollath, J. (1990) [Changes in the lungs and pleura following chemoembolization of liver tumors with mitomycin-lipiodol]. *Rofo* 152, 569-73.
60. Shou, C., Bhardwaj, N., Lam, H. Y. K., Yan, K.-K., Kim, P. M., Snyder, M., & Gerstein, M. B. (2011) Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol* 7, e1001050.
61. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., & Gingeras, T. R. (2012) Landscape of transcription in human cells. *Nature* 489, 101-8.
62. Habegger, L., Sboner, A., Gianoulis, T. A., Rozowsky, J., Agarwal, A., Snyder, M., & Gerstein, M. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27, 281-3.
63. Du, J., Leng, J., Habegger, L., Sboner, A., McDermott, D., & Gerstein, M. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. *PLoS One* 7, e29175.
64. Jee, J., Rozowsky, J., Yip, K. Y., Lochovsky, L., Bjornson, R., Zhong, G., Zhang, Z., Fu, Y., Wang, J., Weng, Z., & Gerstein, M. (2011) ACT: aggregation and correlation toolbox for analyses of genome tracks. *Bioinformatics* 27, 1152-4.
65. Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., Agarwal, A., Auerbach, R., Rozowsky, J., Cheng, C., Kato, M., Miller, D. M., Slack, F., Snyder, M., Waterston, R. H., Reinke, V., & Gerstein, M. B. (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* 21, 276-85.
66.Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., Davis, C. A., Carninci, P., Lassman, T., Gingeras, T. R., Guigó, R., Birney, E., Weng, Z., Snyder, M., & Gerstein, M. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* 22, 1658-67.
67. Cheng, C., Shou, C., Yip, K. Y., & Gerstein, M. B. (2011) Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. *Genome Biol* 12, R111.
68. Cheng, C. & Gerstein, M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 40, 553-68.
69. Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* 11, 533-8.
70. Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E., & Iyer, V. R. (2010) Allele-specific and heritable chromatin signatures in humans. *Hum Mol Genet* 19, R204-9.
71. Ji, H., Li, X., Wang, Q.-f., & Ning, Y. (2013) Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci U S A* 110, 6789-94.

72. Younesy, H., Möller, T., Heravi-Moussavi, A., Cheng, J. B., Costello, J. F., Lorincz, M. C., Karimi, M. M., & Jones, S. J. M. (2013) ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* , .
73. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., & Gerstein, M. B. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* 39, 7058-76.
74. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S., & Kellis, M. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* 23, 800-11.
75. Killela, P. J., Reitman, Z. J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, Jr, L. A., Friedman, A. H., Friedman, H., Gallia, G. L., Giovanella, B. C., Grollman, A. P., He, T.-C., He, Y., Hruban, R. H., Jallo, G. I., Mandahl, N., Meeker, A. K., Mertens, F., Netto, G. J., Rasheed, B. A., Riggins, G. J., Rosenquist, T. A., Schiffman, M., Shih, I.-M., Theodorescu, D., Torbenson, M. S., Velculescu, V. E., Wang, T.-L., Wentzensen, N., Wood, L. D., Zhang, M., McLendon, R. E., Bigner, D. D., Kinzler, K. W., Vogelstein, B., Papadopoulos, N., & Yan, H. (2013) TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A* 110, 6021-6.
76. Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., da Rocha, A. G., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M., Cameselle-Teijeiro, J., Sobrinho-Simões, M., Lima, J., Máximo, V., & Soares, P. (2013) Frequency of TERT promoter mutations in human cancers. *Nat Commun* 4, 2185.
77. Touzet, H. & Varré, J.-S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol* 2, 15.
78. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28, 1045-8.
79. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. (2004) A census of human cancer genes. *Nat Rev Cancer* 4, 177-83.
80. Wagle, N., Berger, M. F., Davis, M. J., Blumenstiel, B., Defelice, M., Pochanard, P., Ducar, M., Van Hummelen, P., Macconaill, L. E., Hahn, W. C., Meyerson, M., Gabriel, S. B., & Garraway, L. A. (2012) High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov* 2, 82-93.
81. Anders, S. & Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol* 11, R106.
82. Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., & Nickerson, D. A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* 7, 250-1.
83. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., & Giegerich, R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22, 500-3.
84. Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27, 91-105.
85. Greenbaum, D., Colangelo, C., Williams, K., & Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 4, 117.
86. Wang, Z., Gerstein, M., & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57-63.
87. Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., Demichelis, F., Rubin, M. A., & Gerstein, M. B. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* 11, R104.
88. Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008) Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Ge-*

nome Res 18, 1865-74.
89. Qian, J., Lin, J., Luscombe, N. M., Yu, H., & Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* 19, 1917-26.
90. Yip, K. Y., Yu, H., Kim, P. M., Schultz, M., & Gerstein, M. (2006) The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* 22, 2968-70.
91. Clarke, D., Bhardwaj, N., & Gerstein, M. B. (2012) Novel insights through the integration of structural and functional genomics data with protein networks. *J Struct Biol* 179, 320-6.
92. Bhardwaj, N., Clarke, D., & Gerstein, M. (2011) Systematic control of protein interactions for systems biology. *Proc Natl Acad Sci U S A* 108, 20279-80.
93. Bhardwaj, N., Abyzov, A., Clarke, D., Shou, C., & Gerstein, M. B. (2011) Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci* 20, 1745-54.
94. Fasolo, J., Sboner, A., Sun, M. G. F., Yu, H., Chen, R., Sharon, D., Kim, P. M., Gerstein, M., & Snyder, M. (2011) Diverse protein kinase interactions identified by protein microarrays reveal novel connections between cellular processes. *Genes Dev* 25, 767-78.
95. Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics* 12, 375.
96. Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O., & 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59-65.
97. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-73.
98. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65.
99. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-8.
100. Zhang, Y., Haraksingh, R., Grubert, F., Abyzov, A., Gerstein, M., Weissman, S., & Urban, A. E. (2013) Child development and structural variation in the human genome. *Child Dev* 84, 34-48.
101. Macarthur, D. G. (2012) Challenges in clinical genomics. *Genome Med* 4, 43.
102. Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A., & Lee, C. (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* 93, 22-6.
103. Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28, 47-55.
104. Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21, 974-84.
105. Abyzov, A. & Gerstein, M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27, 595-603.
106. Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., & Ger-

stein, M. B. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10, R23.
107. Wang, L.-Y., Abyzov, A., Korbel, J. O., Snyder, M., & Gerstein, M. (2009) MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res* 19, 106-17.
108. Zhang, Z. D. & Gerstein, M. B. (2010) Detection of copy number variation from array intensity and sequencing read depth using a stepwise Bayesian model. *BMC Bioinformatics* 11, 539.
109. Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., & Zhu, J. (2013) Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Sci Rep* 3, 2652.
110. Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., & Feolo, M. (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 42, D975-9.
111. International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M. F., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O. M., Joly, Y., Kato, K., Kennedy, K. L., Nicolás, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clément, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P. A., Aburatani, H., Bayés, M., Botwell, D. D. L., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., López-Otín, C., Majumder, P., Marra, M., McPherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., Flicek, P., Getz, G., Guigó, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M., Li, Q., López-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B. F. F., Pearson, J. V., Puente, X. S., Quesada, V., Raphael, B. J., Sander, C., Shibata, T., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Stein, L. D., Guigó, R., Hubbard, T. J., Joly, Y., Jones, S. M., Kasprzyk, A., Lathrop, M., López-Bigas, N., Ouellette, B. F. F., Spellman, P. T., Teague, J. W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K. L., Axton, M., Dyke, S. O. M., Futreal, P. A., Gerhard, D. S., Gunter, C., Guyer, M., Hudson, T. J., McPherson, J. D., Miller, L. J., Ozenberger, B., Shaw, K. M., Kasprzyk, A., Stein, L. D., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cros, A., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D. R. C., Hasel, K. W., Joly, Y., Kaan, T. S. H., Kennedy, K. L., Knoppers, B. M., Lowrance, W. W., Masui, T., Nicolás, P., Rial-Sebbag, E., Rodriguez, L. L., Vergely, C., Yoshida, T., Grimmond, S. M., Biankin, A. V., Bowtell, D. D. L., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. B., Gardiner, B. A., Kench, J. G., Scarpa, A., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., McPherson, J. D., Gallinger, S., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., Chin, L., DePinho, R. A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R. E., Uhlén, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton, M. R., Futreal, P. A., Birney, E., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. A., Martin, S., Reis-Filho, J. S., Richardson, A. L., Sotiriou, C., Stunnenberg, H. G., Thoms, G., van de Vijver, M., van't Veer, L., Calvo, F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J. D., Lathrop, M., Pauporté, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo, F., Bioulac-Sage, P., Clément, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., Zucman-Rossi, J., Lichter, P., Eils, R., Brors, B., Korbel, J. O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifenberger, G., Taylor, M. D., von Kalle, C., Majumder, P. P., Sarin, R., Rao, T. S., Bhan, M. K., Scarpa, A., Pederzoli, P., Lawlor, R. A., Delledonne, M., Bardelli, A., Biankin, A. V.,

Grimmond, S. M., Gress, T., Klimstra, D., Zamboni, G., Shibata, T., Nakamura, Y., Nakagawa, H., Kusada, J., Tsunoda, T., Miyano, S., Aburatani, H., Kato, K., Fujimoto, A., Yoshida, T., Campo, E., López-Otín, C., Estivill, X., Guigó, R., de Sanjosé, S., Piris, M. A., Montserrat, E., González-Díaz, M., Puente, X. S., Jares, P., Valencia, A., Himmelbauer, H., Himmelbaue, H., Quesada, V., Bea, S., Stratton, M. R., Futreal, P. A., Campbell, P. J., Vincent-Salomon, A., Richardson, A. L., Reis-Filho, J. S., van de Vijver, M., Thomas, G., Masson-Jacquemier, J. D., Aparicio, S., Borg, A., Børresen-Dale, A.-L., Caldas, C., Foekens, J. A., Stunnenberg, H. G., van't Veer, L., Easton, D. F., Spellman, P. T., Martin, S., Barker, A. D., Chin, L., Collins, F. S., Compton, C. C., Ferguson, M. L., Gerhard, D. S., Getz, G., Gunter, C., Guttmacher, A., Guyer, M., Hayes, D. N., Lander, E. S., Ozenberger, B., Penny, R., Peterson, J., Sander, C., Shaw, K. M., Speed, T. P., Spellman, P. T., Vockley, J. G., Wheeler, D. A., Wilson, R. K., Hudson, T. J., Chin, L., Knoppers, B. M., Lander, E. S., Lichter, P., Stein, L. D., Stratton, M. R., Anderson, W., Barker, A. D., Bell, C., Bobrow, M., Burke, W., Collins, F. S., Compton, C. C., DePinho, R. A., Easton, D. F., Futreal, P. A., Gerhard, D. S., Green, A. R., Guyer, M., Hamilton, S. R., Hubbard, T. J., Kallioniemi, O. P., Kennedy, K. L., Ley, T. J., Liu, E. T., Lu, Y., Majumder, P., Marra, M., Ozenberger, B., Peterson, J., Schafer, A. J., Spellman, P. T., Stunnenberg, H. G., Wainwright, B. J., Wilson, R. K., & Yang, H. (2010) International network of cancer genome projects. *Nature* 464, 993-8.

112. 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., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., & Getz, G. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-8.

113. Chen, C.-L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., & Thermes, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* 20, 447-57.

114. Schuster-Böckler, B. & Lehner, B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488, 504-7.

115. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, Jr, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., & Garraway, L. A. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603-7.

116. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004) Ultraconserved elements in the human genome. *Science* 304, 1321-5.

117. Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15, 901-13.

118. Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 10, 451-81.

119. Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Mossé, Y. P., Wood, A., Lynch, J. E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E. A., McGrady, P. W., Blakemore, A. I. F., London, W. B., Shaikh, T. H., Bradfield, J., Grant, S. F. A., Li, H., Devoto, M., Rappaport, E. R., Hakonarson, H., & Maris, J. M. (2009) Copy number variation at 1q21.1 as-

sociated with neuroblastoma. *Nature* 459, 987-91.
120. Shlien, A., Tabori, U., Marshall, C. R., Pienkowska, M., Feuk, L., Novokmet, A., Nanda, S., Druker, H., Scherer, S. W., & Malkin, D. (2008) Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A* 105, 11264-9.
121. Bartsch, G., Horninger, W., Klocker, H., Pelzer, A., Bektic, J., Oberaigner, W., Schennach, H., Schäfer, G., Frauscher, F., Boniol, M., Severi, G., Robertson, C., Boyle, P., & Tyrol Prostate Cancer Screening Group (2008) Tyrol Prostate Cancer Demonstration Project: early detection, treatment, outcome, incidence and mortality. *BJU Int* 101, 809-16.
122. Oberaigner, W., Horninger, W., Klocker, H., Schönitzer, D., Stühlinger, W., & Bartsch, G. (2006) Reduction of prostate cancer mortality in Tyrol, Austria, after introduction of prostate-specific antigen testing. *Am J Epidemiol* 164, 376-84.
123. Stankiewicz, P. & Lupski, J. R. (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61, 437-55.
124. Cirulli, E. T. & Goldstein, D. B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11, 415-25.
125. Kirkpatrick, J. (1998) Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *J Insur Med* 30, 204-5.
126. Setlur, S. R., Chen, C. X., Hossain, R. R., Ha, J. S., Van Doren, V. E., Stenzel, B., Steiner, E., Oldridge, D., Kitabayashi, N., Banerjee, S., Chen, J. Y., Schäfer, G., Horninger, W., Lee, C., Rubin, M. A., Klocker, H., & Demichelis, F. (2010) Genetic variation of genes involved in dihydrotestosterone metabolism and the risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev* 19, 229-39.
127. Schaefer, G., Mosquera, J.-M., Ramoner, R., Park, K., Romanel, A., Steiner, E., Horninger, W., Bektic, J., Ladurner-Rennau, M., Rubin, M. A., Demichelis, F., & Klocker, H. (2013) Distinct ERG rearrangement prevalence in prostate cancer: higher frequency in young age and in low PSA prostate cancer. *Prostate Cancer Prostatic Dis* 16, 132-8.
128. Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., & Dermitzakis, E. T. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848-53.
129. Yu, J., Yu, J., Mani, R.-S., Cao, Q., Brenner, C. J., Cao, X., Wang, X., Wu, L., Li, J., Hu, M., Gong, Y., Cheng, H., Laxman, B., Vellaichamy, A., Shankar, S., Li, Y., Dhanasekaran, S. M., Morey, R., Barrette, T., Lonigro, R. J., Tomlins, S. A., Varambally, S., Qin, Z. S., & Chinnaiyan, A. M. (2010) An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* 17, 443-54.
130. Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P.-C., Svensson, M. A., Kitabayashi, N., Moss, B. J., MacDonald, T. Y., Cao, X., Barrette, T., Tewari, A. K., Chee, M. S., Chinnaiyan, A. M., Rickman, D. S., Demichelis, F., Gerstein, M. B., & Rubin, M. A. (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res* 21, 56-67.
131. Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., & Jaenisch, R. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* 153, 910-8.