CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing

Yuchao Jiang[1,2,*], Ioannis N. Anastopoulos[3], Katherine L. Nathanson[3,4], Nancy R. Zhang[5,*]

[1] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.

[2] Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[3] Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[4] Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[5] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

* To whom correspondence should be addressed. Tel: (+1) 215-898-8222; Fax: (+1) 215-898-1280; Email: yuchaoj@email.unc.edu, nzh@wharton.upenn.edu.

Abstract (100 words)

High-throughput DNA sequencing enables detection of copy number variations (CNVs) on the genome-wide scale with finer resolution compared to array-based methods, but suffers from biases and artifacts that lead to false discoveries and low sensitivity. We describe CODEX2, a statistical framework for full-spectrum CNV profiling that is sensitive for variants with both common and rare population frequencies and that is applicable to study designs with and without negative control samples. We demonstrate and evaluate CODEX2 on whole-exome and targeted sequencing data, where biases are the most prominent. CODEX2 outperforms existing methods and, in particular, significantly improves sensitivity for common CNVs.

Key words: copy number variation, normalization, next-generation sequencing, latent factor, negative control.

Background

Copy number variations (CNVs) are large deletions and duplications of segments of the chromosome. CNVs are pervasive in the human genome and play a causal role in diseases such as cancer [1]. In the study of disease, CNVs usually appear in two contexts: *germline* CNVs refer to inherited variants, many of which are polymorphic at the population level [2]; in contrast, *somatic* CNVs, also referred to as copy number aberrations (CNAs), are the copy number changes resulting from somatic mutations, such as those commonly observed in cancer. Germline CNVs can also be described as *common* or *rare* based on their population frequencies. This paper addresses the problem of detection of both germline and somatic CNVs, and, in particular, of improving detection sensitivity for common CNVs in both categories.

With the dramatic growth of sequencing capacity and the accompanying drop in sequencing cost, massively parallel next-generation sequencing offers appealing platforms for genome-wide CNV detection. Whole-genome sequencing (WGS) offers an unbiased genome-wide approach to detect CNVs, while whole-exome sequencing (WES) and targeted sequencing allow the identification of disease-associated variants in coding regions with direct functional interpretation. Despite the rapid technological progress, CNV detection using high-throughput sequencing still faces analytical challenges due to the pervasive biases and artifacts that are introduced during library preparation and sequencing. Proper data normalization is crucial, especially for WES and targeted sequencing, where technical biases are usually magnitudes larger than CNV signal.

For studies in which deep coverage of specific genome regions is desired [3-5], or where the cohort of interest is large, WES and targeted sequencing are often preferred as cheaper alternatives to WGS. For example, the DiscovEHR Collaboration (http://www.discovehrshare.com) has sequenced the exomes of more than 50,000 participants, and the Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org) has generated WES data for 60,706 unrelated individuals taken from multiple disease-specific and population genetic studies. This paper was primarily motivated by the challenges in WES and targeted sequencing data, but the methods developed here have also been integrated into a CNV detection pipeline for WGS data (Zhou et al., https://doi.org/10.1101/172700), where it is shown that the

normalization methods prescribed here allow for more accurate CNV calls using Mendelian concordance as metric.

CNV detection (by WES and targeted sequencing) is primarily based on the detection of local changes in read coverage along the genome. This analysis scheme is based on the simple intuition that regions with copy number gain should have increased read coverage, and regions with copy number loss should have decreased read coverage. However, read coverage depends not just on copy number but also on many other factors, such as GC content [6], mappability [7], and other local sequence characteristics [8]. Therefore, read coverage tends to fluctuate even when there is no CNV, and is especially variable in WES and targeted sequencing data due to the biases and artifacts introduced during the targeting and amplification steps [9-11].

Many methods are available for CNV detection in high throughput sequencing data [9-15]. Despite much progress however [16], significant challenges still remain. Independent benchmark results from multiple studies [9, 11, 13, 17] show that existing methods suffer from low precision and recall rates, especially for the detection of common germline CNV signals. These results are not unexpected, since in WES and targeted sequencing data, and also to an extent in WGS data, the technical background bias for each genomic target varies across samples, leading to false positives and negatives if not properly removed. To remove this technical background, recent methods have relied on low-dimensional linear factor models to capture the background bias [9, 10], which tend to control for false positives. However, these low-dimensional linear factor models tend to remove common CNVs that correlate with the estimated factors. In this paper, we demonstrate that this issue also affects the detection of somatic CNVs in cancer genomes, as CNVs that are recurrent across multiple cancer samples can be accidentally removed in the normalization step. Due to these limitations, current genetic studies using WES and targeted sequencing data have focused mostly on single nucleotide variations and, at best, rare CNVs [18-20].

Herein, we propose CODEX2 for full-spectrum CNV detection in high-throughput sequencing data. In this context, full-spectrum implies the sensitive detection of both common and rare CNVs. CODEX2 can be applied to two scenarios: when there are ‘cases’ and ‘controls’ and the goal is to detect CNVs that are enriched in the case samples

(Figure 1A); and when control samples are lacking and the goal is simply to profile all CNVs in all samples (Figure 1B). We evaluated CODEX2 in three ways. First, we reanalyzed the WES data of the HapMap samples from the 1000 Genomes Project [2], with matched microarrays and experimental validation [21-23] to assess performance. Our results demonstrate that CODEX2 significantly improves both sensitivity and specificity over existing methods, especially for common CNVs. Next, we applied CODEX2 to targeted sequencing data of melanoma cancer cell lines, patient derived xenografts (PDX) and tumor biopsies, and successfully identified recurrent CNVs whose frequencies are highly concordant with those obtained from a separate cohort studied by the Cancer Genome Atlas (TCGA) [1]. Finally, we performed extensive simulations to benchmark existing methods and to explore how key variables, such as population frequency and CNV length, influence performance. CODEX2 is compiled as an open-source R-package available at https://github.com/yuchaojiang/CODEX2.

Results

Methods overview

Figure 1 illustrates the two experimental designs for which CODEX2 can be applied: (i) case-control design with a group of negative control samples, where the goal is to detect CNVs disproportionately present in the 'cases' versus the 'controls' (Figure 1A); and (ii) detection of all CNVs present in all samples design, such as in ExAC (Figure 1B). The key innovation in CODEX2 is the usage of *negative control genome regions* in a genome-wide latent factor model for sample- and position-specific background correction, and the utilization of *negative control samples*, under a case-control design, to further improve background bias estimation under this model. The negative control genome regions defined by CODEX2 are regions that do not harbor *common* CNVs, but that are still allowed to harbor rare CNVs, and can be constructed from existing studies or learned from data.

Figure 2 illustrates how CODEX2 normalization improves upon CODEX and SVD-based normalization methods such as XHMM [10]. For simplicity, and without loss of generality, we represent the background bias by a 1-dimensional latent factor, which we call the "latent batch effect". Against this background, we further assume that there are

two CNVs: Region A, in which the carrier status (the vector indicating whether each sample is a carrier) is correlated with the underlying latent batch, and region B, in which the carrier status is uncorrelated with the latent batch. The signal for these two CNV regions is obscured by the background batch effect in the observed data. In a standard SVD or CODEX normalization, which does not make use of negative controls, all samples and all genomic regions are used in the estimation of the latent background model, resulting in contamination of background estimates by the CNV signal. The contamination is especially severe for CNV A, which is correlated with the one-dimensional background batch. For CODEX2, in scenarios for which negative controls are available, only the negative controls are used to fit the latent factor model, which is then used to predict the background bias of the rest of the samples. In scenarios for which negative control samples are not available but negative control regions are identified, the latent factor model is fitted only using the negative control regions, and then used to "fill in" the background bias for the rest of the regions. In this way, we avoid the contamination of the background estimates by the CNV signal, thus attaining better separation of signal from noise, as can be seen from the histograms of the normalized z-scores. More details are given in Methods.

CODEX2 estimates a separate background factor for each genomic target/region in each sample, which can then be used to normalize the observed coverage and detect CNV regions using the recursive Poisson-likelihood segmentation algorithm in Jiang et al. [11].

Analysis of WES data from the 1000 Genomes Project

We first evaluated CODEX2 by reanalyzing a publicly available WES data set from the 1000 Genomes Project [2], which contains 90 healthy individuals. Forty individuals are Utah residents with northern and western European ancestry; 20 are Japanese from Tokyo; 26 are Yoruba people from Ibadan. Gender is well balanced with 44 males and 46 females. All samples have CNV profiles from three previous microarray studies with experimental validation [21-23]. Please refer to Jiang et al. [11] for more details on the sample input and the microarray studies.

137 To assess performance of CODEX2 and to benchmark against existing methods,
138 we used the experimentally validated CNVs from three previous studies [21-23] as gold
139 standards. Specifically, we adopted a stringent quality control procedure (i.e., the reported
140 CNV must overlap with at least two and at most 20 exons, have less than 5% NA rate
141 across all samples, and have at most three copy number states). These 'gold-standard'
142 CNVs were categorized as common if they are present in more than 10% of the samples,
143 and rare otherwise. Using this 'validation set', we were able to assess the performance
144 of XHMM [10], EXCAVATOR [13], CODEX [11], and CODEX2 to detect common and rare
145 CNVs. Figure 3 shows the precision and recall rates across the four benchmarked
146 methods, using the CNVs validated by each of the three studies [21-23] as ground truth.
147 The grey lines are the contours of the F-measure, defined as the harmonic mean of
148 precision and recall. XHMM, which is designed for detection of rare CNVs, lacks
149 sensitivity to common CNVs, so does EXCAVATOR. CODEX detects proportionately
150 more common CNVs but still lacks sensitivity. The best of the three existing methods,
151 CODEX, achieves less than 40% recall rate for common CNVs. CODEX2 achieves a
152 recall rate of 87.6%, 57.2%, and 83.1% in the three validation sets, respectively, while
153 simultaneously making substantial improvements to specificity. Across all three validation
154 sets, CODEX2 achieves good performance for both common and rare CNVs, with
155 significant improvement in precision and recall for common CNV detection.
156
157 ***Analysis of targeted sequencing data of melanoma cases and controls***
158 We further applied CODEX2 to a cohort study of melanoma from Garman et al. (in press
159 at Cell Reports) including 334 cases (untreated human melanoma cell lines, patient-
160 derived xenografts, and tumors) and 16 controls. Samples were sequenced on a custom
161 capture panel of 108 genes previously implicated in tumorigenesis of melanoma. For
162 almost all tumor suppressor genes, the entire gene region (exons and introns) were
163 sequenced to facilitate CNV calling; for oncogenes, only exons were sequenced. For
164 cases where the full gene is captured and sequenced, we separated the gene region into
165 consecutive windows of 500 base pairs. This resulted in a panel of 1398 targets across
166 350 samples.

We applied CODEX2 to this data set and compare to CODEX. The number of Poisson latent factors in the background model is determined by the Bayesian information criterion (BIC) for both programs. The use of negative controls in estimating the background model allowed CODEX2 to be more robust to model tuning. For CODEX2, the number of latent factors had minimal effect on normalization and more generally on CNV detection, as only the normal samples were used to estimate the bias coefficient for each exon (Supplementary Table 1). In comparison, for CODEX, the number of CNV events decreased as the number of latent factors increased (Supplementary Table 1). Since the 108 genes are sparsely scattered across the genome, segmentation is performed within each gene separately. Furthermore, due to clonal heterogeneity and normal cell contamination, copy numbers may not be integers, and are assumed to be continuous and fractional to represent attenuated mean estimates of the genome mixtures. We categorize a CNA event to be high gain, gain, diploid, one-copy deletion, and two-copy deletion, if the profiled copy number is above 3.3, between 2.3 and 3.3, between 1.7 and 2.3, between 0.7 and 1.7, and below 0.7, respectively. Figure 4 shows the heatmaps of the segmentation results by CODEX and CODEX2. Each row corresponds to a sample, with the first 16 samples towards the bottom corresponding to the normal controls; each column corresponds to a target in the gene panel. In melanoma, somatic deletions of tumor suppressors (e.g., *PTEN*) and duplications of oncogenes (e.g., *BRAF*) are known to have high incidence rates [1]. From visual inspection of the heatmaps in Figure 4, we see that CODEX2 successfully retains these expected recurrent deletions and duplications, while CODEX, which does not make use of the negative control samples in fitting the background model, misinterprets the recurrent signals as a background latent factor.

To rigorously evaluate CODEX2's accuracy on this data, we compared the frequencies of the profiled gains and losses, that is, the proportion of samples in which a call is made, with frequencies from an independent melanoma patient cohort in TCGA [1]. Specifically, for each gene target, we plotted in Figure 5 the proportion of samples carrying a deletion (or duplication) in TCGA, versus this corresponding proportion in our current data set. CODEX2 achieves much higher concordance with TCGA results, with overall correlation across genes reaching 0.842 for deletions and 0.853 for gains, as

compared to CODEX (correlation = 0.52 for deletions and 0.049 for gains). CODEX2 detects in these cell line samples a higher frequency of *BRAF* amplification and *CDKN2A* loss, as compared to the frozen-tissue derived TCGA results, which is not surprising due to the relative *in vitro* growth advantage of cells carrying these mutations. Based on the results by CODEX2, Garman et al. (in press at Cell Reports) further separated the cohort based on cancer subtypes and clinicopathological characteristics (responses to targeted and/or immunotherapy) and investigated the differences in mutational profiles between groups. The accurate profiling of CNVs in this data set enables unbiased downstream analysis.

Performance assessment via spike-in studies

To understand how variables such as CNV length and population frequency affected the sensitivity of CODEX, CODEX2, and methods based on singular value decomposition (SVD, such as XHMM and CoNIFER), we conducted *in silico* spike-in studies. We started with the exon read depth data from chromosomes 16 to 22 in the 90 samples we analyzed from the 1000 Genomes Project, and applied filtering to remove putative existing CNVs. We then added, to the background count matrix under the null, heterozygous CNV signals of varying length, frequency, and degree of correlation with the first latent factor in the background model. See details of simulation setup in Methods.

As an illustration, Figure 6 shows a small subset of the CNV regions in the spike-in data with the ground truth, the post-normalization heatmap, and the CNV assignments across multiple methods, with the "null" regions containing no CNVs removed for easier visualization. The histograms in Figure 6 show the distribution of the normalized z-scores, with exons that harbor CNVs in red and exons within diploid regions in grey. We see that CODEX2 achieves clear separation of the deletions and diploids with the distributions centered on the expected values, $\log(1/2)$. The segmentations by CODEX and XHMM contain false negatives as well as false positives, especially in regions where common CNV signals reside, whereas improved normalization by CODEX2 allows almost perfect segmentation. Supplementary Figure 3 further shows the true latent background parameters (estimated using raw read depth without spike-in, which represents the null model that would be unobservable in a real data scenario) and the estimated parameters

obtained by CODEX and CODEX2 on the "observed data" containing added CNVs. Our results show that while the sample loadings $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3\}$ are consistent between CODEX, CODEX2, and the ground truth, the exon-specific factors $\{\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3\}$ estimated by CODEX are biased due to the inclusion of the mean CNV signals, reflecting the same trend in Supplementary Figure 2. CODEX2 corrects this bias through the use of the mixture model.

We systematically compare the performance of CODEX2 against existing methods by spiking in CNVs of length 5, 10, 20, and 40 exons with population frequency $p \in \{5\%, 10\%, \ldots, 95\%\}$, repeating each simulation run 20 times. The precision and recall rates achieved by each method are shown in Figure 7. The plots show how two variables: population frequency and degree of correlation with batch effect, impact the accuracy of methods. The results show that CODEX and SVD-based methods are sensitive to both variables, while CODEX2 maintains high accuracy across all frequencies and all degrees of correlation. CODEX2 has nearly perfect performance, whereas CODEX and SVD-based methods suffer from low sensitivity and specificity, especially for common CNVs with frequencies around 50%. We also investigated the effect of CNV length on performance and demonstrated that CODEX and SVD-based methods have lower sensitivity and specificity for longer CNVs, as compared to CODEX2 (Supplementary Figure 4).

We further studied the relationship between CNV carrier status and batch effects—44 and 46 samples are sequenced by the Baylor College of Medicine and the Washington University Genome Sequencing Center respectively—and show that when they are highly correlated, CODEX suffers from low sensitivity, while CODEX2 is able to recover the CNV signals from the batch effect (Figure 7). Detailed results from an example run are shown in Supplementary Figure 5, where the spiked-in CNVs are highly correlated with the batch effect (i.e., most of the carriers are from one of the sequencing centers). CODEX estimates the latent factors assuming all exons are null, with the fitted regression line lying between the carriers and non-carriers, resulting in low sensitivity for true deletions in carriers and false positives as duplications in non-carriers. CODEX2, on the other hand, estimates the exon-specific latent factors in common CNV regions using the proposed EM algorithm and successfully separates the carriers from the non-carriers, which leads

to clean down-stream segmentation. Here we show that normalization is of first-order effect.

Methods

Background and overview

Multiple efforts have been made to recover CNV signals from experimental noise. VarScan2 [12], ExomeDepth [14], and ExomeCNV [15] control for baseline fluctuations in read coverage by relying on matched normal samples or building an optimized reference set. EXCAVATOR [13] adopts a median normalization approach for bias removal. It was soon realized that the magnitudes of the various sources of bias are sample-specific, and thus cannot be completely removed by normalizing to control samples or reference sets. This realization motivated the development of CoNIFER [9] and XHMM [10] which adopt SVD to estimate sample-specific backgrounds that can be more effective. SVD, however, is designed for capturing linear biases in continuous-valued observations. GC content has been shown to have a sample-specific, non-linear bias in sequencing data. Furthermore, read counts do not fit well by Gaussian models, even after transformation, due to their fluctuation over a very wide range. Our previous work, CODEX [11], adopts a Poisson latent factor model for count-based sequencing data and estimates a sample-specific background for each genomic position that incorporates nonlinear biases due to GC content, target-specific capture and amplification efficiency, and low-dimensional latent systemic artifacts.

We will start by giving an overview of the SVD-based methods by CoNIFER [9] and XHMM [10] and the Poisson latent factor model by CODEX [11]. We will discuss the limitations of existing models and the reason why they lack sensitivity for common CNVs. We will then describe the model for CODEX2, leaving algorithmic details to the Supplements.

Review and reevaluation of existing methods

Denote $\boldsymbol{Y}$ as a $n \times m$ matrix of raw read depth, where Y_{ij} corresponds to the read depth for exon $i \in \{1, \dots, n\}$ in sample $j \in \{1, \dots, m\}$. SVD-based methods CoNIFER [9] and XHMM [10] remove the strongest K SVD components,

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{U}_K \boldsymbol{D}_K \boldsymbol{V}_K^T,$$

where $\boldsymbol{U}_K$ and $\boldsymbol{V}_K$ correspond to the top K left and right singular vectors, respectively, and $\boldsymbol{D}_K$ corresponds to the diagonal matrix of the K largest singular values. Each column of $\boldsymbol{Y}^*$ is column-standardized. The genome is then segmented by hidden Markov model (HMM) into 'diploid', 'deletion', or 'duplication' states.

This paper extends CODEX [11], which improves upon SVD-based approaches in several ways: CODEX adopts a Poisson model that more accurately models count data, and importantly, it explicitly models observable and measurable sources of bias, such as GC content and exon length, in addition to unmeasurable biases, due to unanticipated experimental variables, in the form of latent factors. In particular, GC content bias exhibits nonlinear patterns of variation across samples [6, 11], and thus CODEX uses a non-linear sample-specific function instead of a low-rank linear factor to capture this bias. Since an understanding of the CODEX model is integral to our ensuing discussion, we review it here. CODEX assumes that

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

$$\lambda_{ij} = N_j f_j(GC_i) \beta_i \exp\left(\sum_{k=1}^{K} g_{ik} h_{jk}\right),$$

where N_j is the total number of reads in sample j, $f_j(\cdot)$ is the GC content bias function for sample j, GC_i is the GC percentage of exon i, β_i is an exon-specific factor capturing multiplicative effects due to features such as exon length, and $\boldsymbol{g}_k = \{g_{1k}, \dots, g_{nk}\}$, $\boldsymbol{h}_k = \{h_{1k}, \dots, h_{mk}\}$ are the kth ($1 \leq k \leq K$) Poisson latent factors. The Poisson latent factors form a low-dimensional background model to capture unanticipated experimental variables, similar to the singular vectors in SVD-based normalization models [9, 10]. CODEX uses maximum likelihood to estimate the parameters of the null model, where $\boldsymbol{g}$ and $\boldsymbol{h}$ are assumed to be orthogonal to ensure identifiability (see Supplementary Algorithm 1 for details). For the remaining details on model selection, parameter estimation, and genome segmentation, see Jiang et al. [11].

These SVD- and factor-based methods, so far, all lack sensitivity for common CNVs. This is because common CNVs bias the estimation of the low-dimensional linear factors ($\boldsymbol{g}_k$ and $\boldsymbol{h}_k$ in CODEX's model and $\boldsymbol{U}_k$, $\boldsymbol{V}_k$ in SVD-based algorithms), which, in CNV regions, capture part of the CNV signal. Therefore, subsequent removal of these

factors also removes the signal for the CNVs. Supplementary Figure 2 shows a toy example illustrating this issue, focusing on one exon i^* and assuming that $K = 1$. For this exon, CNV signals are *in silico* added at six population frequencies, creating six scenarios. For each scenario, simulated read counts $\boldsymbol{Y}_i = \{Y_{i1}, \dots, Y_{im}\}$ are plotted against the sample factors $\boldsymbol{h}_1 = \{h_{11}, \dots, h_{m1}\}$. The samples that carry the CNV are shown in red, the rest are shown in black. In each iteration of CODEX, Poisson log-linear regression is performed for $\boldsymbol{Y}_i$ on $\boldsymbol{h}_1$ to get g_{i1}, the 'loading' for exon i. CODEX estimates g_{i1} assuming that the great majority of the samples are 'nulls', that is, they don't carry CNVs, resulting in estimated background values shown in green. Ideally, the background curve should be fit to the control samples, the true 'nulls', as shown in blue. Yet in real data we don't know which are the true carriers of this CNV, and the background curve is thus contaminated with signal, lying between the carriers and the non-carriers. The higher the incident rate of the CNV, the higher the contamination of the background by signal, and the closer the fitted green curve to the carriers rather than the non-carriers. This, as a result, leads to low sensitivity in detecting the true CNV carriers. The effect of population frequency on detection sensitivity by CODEX and SVD-based algorithms is also explored by simulation shown in Figure 6 and Figure 7.

CODEX2 model and full-spectrum CNV detection

Figure 1 gives an overview of the two scenarios of CODEX2. To describe CODEX2, we first consider the simpler scenario where negative control samples are available. Without loss of generality, we consider a cohort comprised of both tumor and normal samples, the normal samples may not be matched to the tumor samples. Our goal is to identify CNVs present in the cancer samples but not in the normal samples. Duplications of oncogenes and deletions of tumor suppressors are commonly seen in cancer samples and have been reported to be associated with cancer. It is therefore crucial to detect somatic CNVs recurrent only in cancer samples with high sensitivity and specificity. Denote $\boldsymbol{J}_c$ as the set of indices of the normal samples, which serve as negative control. Assuming that the normal samples are copy-number-neutral, the negative control samples are used by CODEX2 to estimate the exon-specific bias $\boldsymbol{\beta}$ and latent factors $\boldsymbol{g}$ to avoid attenuating and removing common CNV signals. Poisson regression is then to each cancer sample

352 to obtain the sample loadings $\boldsymbol{h}$. The sample-specific background values are then
353 computed using $\boldsymbol{\beta}, \boldsymbol{g}, \boldsymbol{h}$ and used in Poisson likelihood segmentation to identify CNVs.
354 Refer to Supplementary Algorithm 2 for the detailed estimation procedure.

355 Now let's consider the scenario where negative control samples are not available.
356 We denote $\boldsymbol{I}^*$ as a set of indices of the exons that harbor highly recurrent CNVs, the
357 compliment of which are the indices of exons within the negative control regions. The set
358 $\boldsymbol{I}^*$ can be obtained based on prior knowledge (e.g., common deletions in tumor
359 suppressors), from existing database (e.g., the Database of Genomic Variants), or
360 empirically from a first-pass CODEX run. That is, if an exon lies within a common CNV
361 region, CODEX will return a high standard deviation of the normalized z-scores across all
362 samples for this exon – Supplementary Figure 2 shows that for a common CNV, the
363 estimated null will be biased towards the alternative, which has high incidence rate. Figure
364 1B shows an example on identifying germline CNVs from a population of samples (e.g.,
365 healthy individuals from the 1000 Genomes Project). For step 1 in Supplementary
366 Algorithm 2, we no longer have a set of controls samples to directly estimate the exon-
367 specific parameters. We propose an expectation-maximization (EM) algorithm embedded
368 in our iterative parameter estimation procedure, where the missing data is the carrier
369 status of the samples. Specifically, for each exon $i \in \boldsymbol{I}^*$,

370 $$Y_{i^*j} \sim \text{Poisson}(\lambda_{i^*j}),$$

371 $$\log(\lambda_{i^*j}) = \log(N_j) + \log(f_j(GC_{i^*})) + \log(\beta_{i^*}) + \sum_{k=1}^{K} g_{i^*k} h_{jk} + \alpha_{i^*j},$$

372 where

373 $$\alpha_{i^*j} = \begin{cases} \mu_{i^*} & \text{with probability } \pi_{i^*}, \\ 0 & \text{with probability } 1 - \pi_{i^*}. \end{cases}$$

374 π_{i^*} is the incident rate for the CNV that span exon i^*, and μ_{i^*} is the deviation from the null
375 on the log scale, which can be either pre-fixed (e.g., $\log(1/2)$ for heterozygous deletion)
376 or estimated by CODEX2. For simplicity, here we show the case where there is only one
377 type of CNV event within the carriers. This can be easily extended to multiple subgroups
378 with a model selection metric, which is enabled in the CODEX2 package. $\boldsymbol{N}$, $\boldsymbol{f}(\boldsymbol{GC})$, and
379 $\boldsymbol{h}$ can be using negative control regions shown in Supplementary Algorithm 3 step 1 and

380 2. We adopt an EM algorithm to estimate the unknown parameters $\Theta =$
381 $\{\beta_{i^*}, g_{i^*1}, \dots, g_{i^*K}, \pi_{i^*}, \mu_{i^*} | i^* \in \boldsymbol{I}^*\}$ with missing data:

382 $$Z_{i^*j} = \begin{cases} 1 & \text{if sample } j \text{ has CNV spanning exon } i^*, \\ 0 & \text{otherwise.} \end{cases}$$

383 Refer to Supplementary Algorithm 3 for implementation details.

384 CODEX2 offers the choice of Akaike information criterion (AIC), Bayesian
385 information criterion (BIC), and variance reduction to determine the optimal number of
386 latent factors. In CODEX and SVD-based methods (XHMM and CoNIFER), the number
387 of latent factors is a critical model tuning parameter that affects normalization and
388 segmentation results. It is often not clear whether to use the AIC, BIC, or simple visual
389 examination of the scree plot, and this arbitrariness plagues all methods that rely on SVD,
390 PCA, or factor models. By using negative controls to guide the estimation of the
391 background model, CODEX2 is less sensitive to the number of latent factors
392 (Supplementary Table 1), thus giving results that are easier to reproduce.

393 For segmentation, CODEX2 adopts the same Poisson likelihood ratio based
394 approach as CODEX. Please refer to Jiang et al. [11] for details. For targeted sequencing,
395 where a smaller pre-selected panel of targets are sequenced, the normalization model is
396 exactly the same; the segmentation is performed within each gene separately. For WGS,
397 refer to Zhou et al. (https://doi.org/10.1101/172700) for details.

398

399 ***Simulation setup***

400 We start with the exonic read depth data of chromosomes 16 to 22 from the 90 samples
401 we analyzed from the 1000 Genomes Project, and apply filtering to remove putative
402 existing CNVs. Specifically, the filtering step removes all exons that: (i) are called to
403 harbor CNV events by XHMM, EXCAVATOR, CODEX, or CODEX2, (ii) overlap with
404 duplication and deletion reported in the Database of Genomic Variants, (iii) don't pass
405 quality control procedure by CODEX (median coverage between 40 and 4000, length
406 between 30 and 2000 base pairs, mappability less than 0.95, GC content between 30%
407 and 70%), (iv) have standard deviation of normalized z-scores across samples above 0.3,
408 or maximum of normalized z-scores above 0.8, or minimum normalized z-scores below -
409 0.8 across all samples. This way we are left with 4035 'null' exons that are CNV-free
410 across 90 samples. Treating this filtered count matrix as background, we fit the

411 background model of CODEX, with the estimated parameters as ground truth. The
412 optimal number of latent factors is 3 by AIC, BIC and variance reduction (Supplementary
413 Figure 3) and is kept the same for subsequent analysis for CODEX2, CODEX, and SVD-
414 based method. We then add, to this background count matrix, heterozygous CNV signals
415 of varying length, frequency, and degree of correlation with the first latent factor in the
416 background model. In more detail, we spike in heterozygous deletions, of varying lengths
417 and population frequencies, by reducing the raw depth of coverage for exons spanned by
418 the CNV from y to $y \times c/2$, where c is sampled from a normal distribution with mean 1 and
419 standard deviation 0.1. Note that heterozygous deletions with frequency p in the
420 population have exactly the same detection accuracy as duplications with frequency $1 -$
421 p, since all copy number events are defined in reference to a population average.
422

423 **Discussion and conclusions**

424 A limitation shared by all existing CNV detection methods, highlighted by multiple
425 independent benchmarking studies, is the lack of sensitivity for common variants.
426 Similarly, in our experience applying CODEX [11] to WES and targeted sequencing of
427 tumor samples, we easily detect sporadic aberrations but miss highly recurrent
428 aberrations. To meet the widespread demand for improved CNV detection, we develop
429 in this paper a new method, CODEX2, to remove technical noise and improve CNV signal-
430 to-noise ratio for all sequencing platforms including WES and targeted sequencing.
431 CODEX2 builds on our existing method CODEX [11] with a significant improvement of
432 sensitivity for common variants, thus allowing full-spectrum CNV detection. CODEX2 can
433 be applied in either the setting of a case-control analysis in which the goal is to detect
434 CNVs that are enriched in the case samples, or when the goal is simply to profile all CNVs.
435 We have benchmarked CODEX2 extensively against existing methods. In the first
436 evaluation, we reanalyzed WES data of the HapMap samples from the 1000 Genomes
437 Project, for which a set of experimentally validated CNV calls from microarrays and other
438 sources could be used to assess performance. Our results demonstrate that CODEX2
439 has markedly improved sensitivity and specificity over existing methods. The
440 improvement for calling of common variants is the most substantial, from 40% recall rate
441 to >80% recall rate, in two out of the three validation sets used, while simultaneously

442 improving precision from 60% to 90%. In the second evaluation, we applied CODEX2 to
443 targeted sequencing data of melanoma cell lines, PDX, and tumor samples, in which
444 CODEX2 detects CNVs with recurrence rates that are highly concordant with those
445 obtained from TCGA. Finally, we performed extensive simulations benchmarking existing
446 methods and elucidated how key variables, such as population frequency, influence
447 detection sensitivity. Together, these results establish the improved accuracy of CODEX2
448 over existing state-of-the-art approaches, and the utility of the software under varying
449 study designs.

450 Under a different context, for the detection of differential expression in RNA
451 sequencing data, Risso et al. [24] proposed the normalization method "removing
452 unwanted variation" (RUV) which is based on a factor model that relies on sets of control
453 genes or samples for estimation. CODEX2 resembles RUV but has distinguishing
454 features. First, RUV is designed to be used in a case-control setting. CODEX2 can be
455 applied with and without negative control samples. Second, when estimating the latent
456 factors that correspond to the control genes or samples, RUV adopts SVD with Gaussian
457 assumption, whereas CODEX2 uses Poisson generalized linear modeling to achieve
458 better fit for low counts. Furthermore, CODEX2 directly models the GC content bias,
459 which cannot be fully captured by a linear principal component [6, 11], as well as library
460 size and exon length, all of which can be directly quantified.

461 CODEX2 normalizes the read depth data for CNV detection via a Poisson latent
462 factor model, which can be well adapted to other settings within the genomics domain.
463 Lee et al. [25] applies a similar Poisson factor approach to the non-normalized microRNA
464 sequencing data. Chen et al. [26] estimates allele-specific copy number under tumor-
465 normal setting using the Poisson latent factor model to remove biases and artifacts that
466 cannot be fully captured by comparing to the normal.

467 Herein, we have shown the results of applying CODEX2 to WES and targeted
468 sequencing data. For WGS, user-defined consecutive bins can be treated as 'targets' in
469 the WES setting, with normalization and segmentation procedures carried out in the same
470 fashion. CODEX2 has been integrated into a pipeline to detect CNV using both
471 sequencing (WGS, WES, and targeted) and microarray data (Zhou et al.,
472 https://doi.org/10.1101/172700). Sequencing capacity has increased exponentially over

473 the past few years. This tremendous amount of data opens up great opportunities and
474 yet at the same time precautions should be made with regard to running efficiency and
475 method scalability. CODEX2 processes each chromosome within each batch separately
476 and can thus be highly parallelized. With increasing sequencing capacity, and increasing
477 need to profile CNVs as a non-negligible source of genetic variation, we believe that
478 CODEX2 can be a useful tool for the genetics and genomics community.
479
480 **Abbreviations**
481 Copy number variation: CNV, whole-genome sequencing: WGS, whole-exome
482 sequencing: WES, singular value decomposition (SVD), the Cancer Genome Atlas
483 (TCGA), Akaike information criterion (AIC), Bayesian information criterion (BIC),
484 expectation-maximization (EM), remove unwanted variation (RUV).
485
486 **Declarations**
487 ***Ethics Approval and consent to participate***
488 Not applicable.
489
490 ***Consent for publication***
491 Not applicable.
492
493 ***Availability of data and material***
494 CODEX2 is an open-source R package available at
495 https://github.com/yuchaojiang/CODEX2 with license GPL-2.0.
496
497 ***Competing interests***
498 The authors declare no conflict of interest.
499
500 ***Funding***
501 This work was supported by the National Institutes of Health (NIH) grant R01 HG006137
502 to NRZ and P01 CA142538 to YJ. Sequencing data generation from melanoma samples

was supported by NIH grants P50 CA174523 (SPORE on Skin Cancer) and P01 CA 114046 to Meenhard Herlyn.

Authors' contributions

NRZ initiated and envisioned the study. YJ and NRZ formulated the model. YJ developed and implemented the algorithm. YJ, KLN, and NRZ conducted the analysis. YJ and NRZ wrote the manuscript, which was edited by KLN. All authors read and approved the final manuscript.

Acknowledgements

We thank Bradley Garman and Bradley Wubbenhorst for providing the case-control melanoma dataset from samples collected and studied at the Wistar Institute in the laboratory of Dr. Meenhard Herlyn, led by Dr. Clemens Krepler. We also thank Xiaoqi Geng, Xiangwei Zhong, and Dr. Jingshu Wang for helpful comments and discussions.

References

1. Cancer Genome Atlas N: **Genomic Classification of Cutaneous Melanoma.** *Cell* 2015, **161:**1681-1696.
2. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526:**75-81.
3. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, Kotsopoulos SK, Pond S, Crain B, Chee MS, Messer K, et al: **Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing.** *Genome Biol* 2011, **12:**R124.
4. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, Ducar M, Van Hummelen P, Macconaill LE, Hahn WC, et al: **High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing.** *Cancer Discov* 2012, **2:**82-93.
5. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, et al: **Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders.** *Science* 2012, **338:**1619-1622.
6. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Res* 2012, **40:**e72.
7. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nat Rev Genet* 2014, **15:**121-132.
8. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A: **Statistical challenges associated with detecting copy number variations with next-generation sequencing.** *Bioinformatics* 2012, **28:**2711-2718.

9. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Project NES, Quinlan AR, Nickerson DA, Eichler EE: **Copy number variation detection and genotyping from exome sequence data.** *Genome Res* 2012, **22:**1525-1532.
10. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, et al: **Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth.** *Am J Hum Genet* 2012, **91:**597-607.
11. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR: **CODEX: a normalization and copy number variation detection method for whole exome sequencing.** *Nucleic Acids Res* 2015, **43:**e39.
12. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res* 2012, **22:**568-576.
13. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, et al: **EXCAVATOR: detecting copy number variants from whole-exome sequencing data.** *Genome Biol* 2013, **14:**R120.
14. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, et al: **A robust model for read count data in exome sequencing experiments and implications for copy number variant calling.** *Bioinformatics* 2012, **28:**2747-2754.
15. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF: **Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV.** *Bioinformatics* 2011, **27:**2648-2654.
16. Tattini L, D'Aurizio R, Magi A: **Detection of Genomic Structural Variants from Next-Generation Sequencing Data.** *Front Bioeng Biotechnol* 2015, **3:**92.
17. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A: **Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2.** *Nucleic Acids Res* 2016, **44:**e154.
18. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, et al: **De novo mutations revealed by whole-exome sequencing are strongly associated with autism.** *Nature* 2012, **485:**237-241.
19. Shi H, Moriceau G, Kong X, Lee MK, Lee H, Koya RC, Ng C, Chodon T, Scolyer RA, Dahlman KB, et al: **Melanoma whole-exome sequencing identifies (V600E)B-RAF amplification-mediated acquired B-RAF inhibitor resistance.** *Nat Commun* 2012, **3:**724.
20. Witkiewicz AK, McMillan EA, Balaji U, Baek G, Lin WC, Mansour J, Mollaee M, Wagner KU, Koduru P, Yopp A, et al: **Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets.** *Nat Commun* 2015, **6:**6744.
21. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, et al: **Integrated detection and population-genetic analysis of SNPs and copy number variation.** *Nat Genet* 2008, **40:**1166-1174.
22. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2010, **464:**704-712.
23. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467:**52-58.
24. Risso D, Ngai J, Speed TP, Dudoit S: **Normalization of RNA-seq data using factor analysis of control genes or samples.** *Nat Biotechnol* 2014, **32:**896-902.
25. Lee S, Chugh PE, Shen H, Eberle R, Dittmer DP: **Poisson factor models with applications to non-normalized microRNA profiling.** *Bioinformatics* 2013, **29:**1105-1111.

587 26. Chen H, Jiang Y, Maxwell KN, Nathanson KL, Zhang N: **Allele-specific copy number estimation by**
588 **whole exome sequencing.** *The Annals of Applied Statistics* 2017, **11:**1169-1192.

589

590 **Figure legends**

591 **Figure 1. CODEX2 can be applied under two experimental designs to identify**
592 **common and rare, germline and somatic CNVs.** (A) CODEX2 with negative control
593 samples. The goal is to recover both rare and common CNV signals present
594 disproportionately in the cases versus the controls. (B) CODEX2 without negative control
595 samples. The goal is to identify all CNVs in all samples, e.g., to detect germline CNVs in
596 healthy individuals without prior knowledge of disease status.

597

598 **Figure 2. Outline of analysis framework with benchmark against SVD-based**
599 **methods.** Results are based on *in silico* simulations where the ground truth with one-
600 dimensional latent factor is known. One CNV signal is correlated with batch effect; the
601 other is not. CODEX2 utilizes negative control samples and negative control regions and
602 outperforms SVD-based methods.

603

604 **Figure 3. Assessment of CNV calls on WES data from the 1000 Genomes Project**
605 **by array-based methods.** CNV calls by XHMM, EXCAVATOR, CODEX and CODEX2
606 are validated against genotyping calls from (A) HapMap 3 Consortium [23], (B) Conrad et
607 al. [22], and (C) McCarroll et al. [21]. Grey contours show F-measures as harmonic
608 means of precision and recall rates. CODEX2 has significantly improved precision and
609 recall for common CNVs and has the highest F-measure among all methods.

610

611 **Figure 4. Heatmap of CODEX and CODEX2 normalization/segmentation results for**
612 **the melanoma cohort.** Each column is one target in the gene panel; each row is one
613 sample, with the first 16 towards the bottom in the heatmap being normal. Profiled CNVs
614 are categorized as high gain, gain, diploid (null), one-copy loss, and two-copy loss based
615 on the estimated copy numbers. Only part of the oncogenes and tumor suppressors with
616 greater than 30 targets are shown.

617

618 **Figure 5. Assessment of profiled CNVs in the melanoma cohort with comparison**
619 **to TCGA.** CNVs are separated by states: losses on the left panels and gains on the right
620 panel. Each dot corresponds to one gene in the targeted sequencing panel. CNV
621 frequencies detected by (A) CODEX and (B) CODEX2 from the melanoma cohort is
622 compared to the TCGA cohort with CODEX2 having drastically higher correlation.
623
624 **Figure 6. Performance assessment by spike-in experiment**. CNV signals are *in silico*
625 spiked in with increasing population frequencies. (A) Ground truth, and normalization
626 results by CODEX, CODEX2, and SVD based method. (B) Histogram of normalized $\boldsymbol{z}$-
627 scores as $\log(\boldsymbol{Y}/\hat{\boldsymbol{\lambda}})$ in CNV (red) and diploid (grey) regions. The $\boldsymbol{z}$-scores returned by
628 XHMM are on a different scale with a much wider range and significant overlap between
629 the null and non-null. (C) Segmentation results where CODEX2 has the highest
630 concordance with the ground truth among the methods being compared.
631
632 **Figure 7. Assessment of precision and recall rates with different CNV frequencies**
633 **and correlations between CNV state and batch effect**. CODEX2 has nearly perfect
634 performance compared to CODEX and XHMM. The latter two suffer from low recall and
635 precision rates, especially for common CNV signals that are highly correlated with the
636 batch effects.

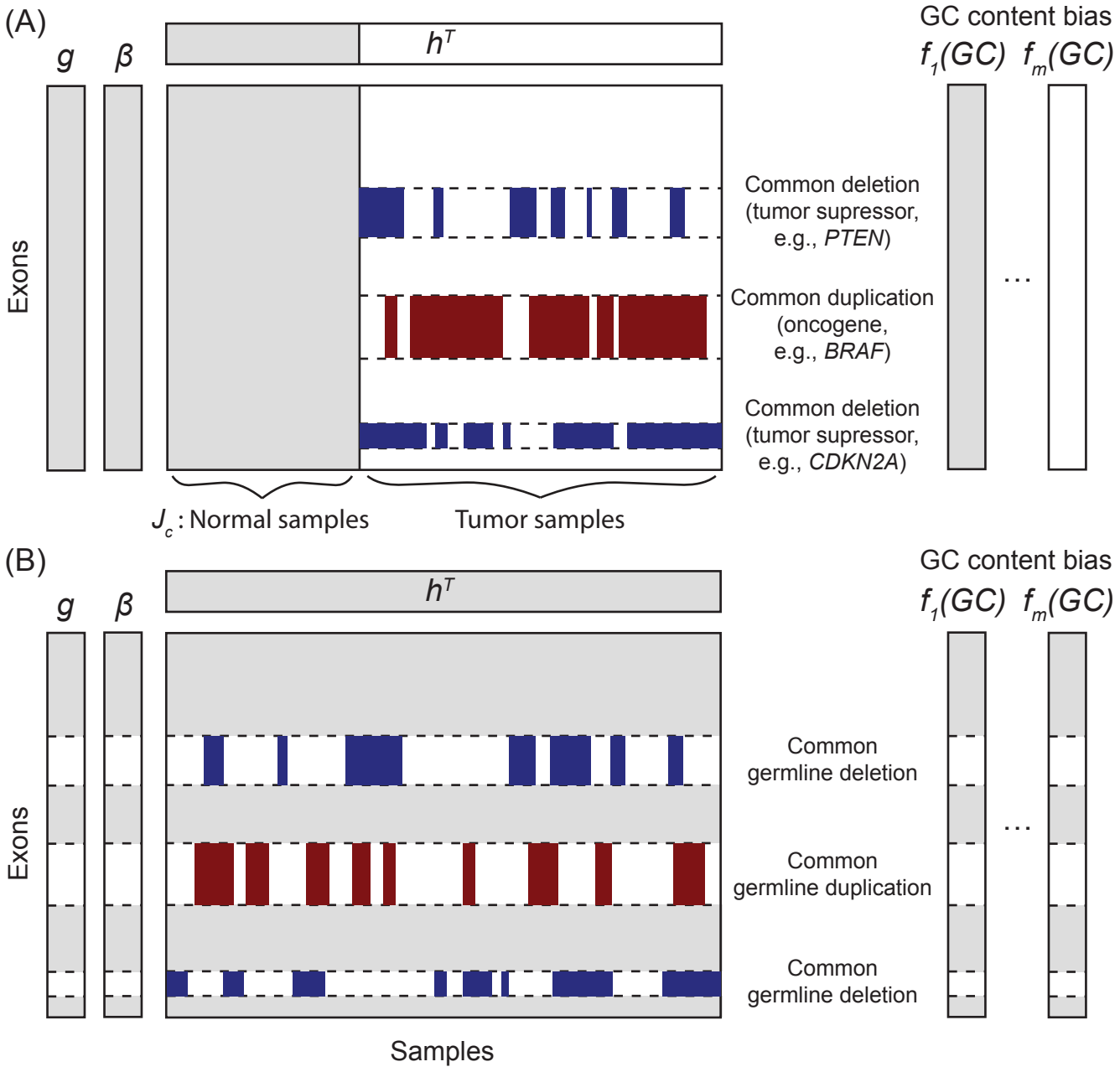

(A)
g
β
h^T
GC content bias
$f_1(GC)$ $f_m(GC)$
Exons
Common deletion
(tumor supressor,
e.g., PTEN)
Common duplication
(oncogene,
e.g., BRAF)
Common deletion
(tumor supressor,
e.g., CDKN2A)
…
J_c : Normal samples
Tumor samples
(B)
g
β
h^T
GC content bias
$f_1(GC)$ $f_m(GC)$
Exons
Common
germline deletion
Common
germline duplication
Common
germline deletion
…
Samples

Figure 1

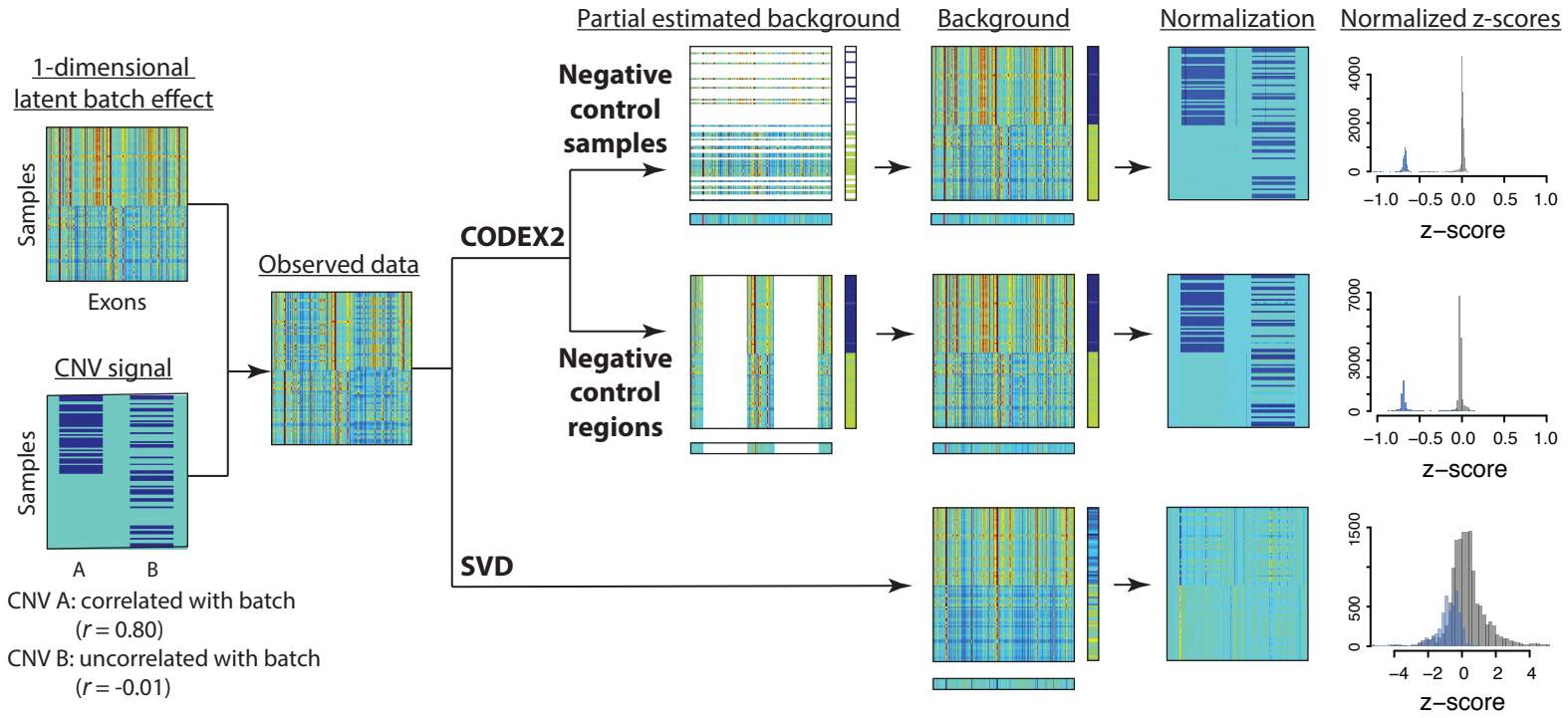

1-dimensional latent batch effect
Samples
Exons
CNV signal
Samples
A
B
CNV A: correlated with batch
(r = 0.80)
CNV B: uncorrelated with batch
(r = -0.01)
Observed data
CODEX2
SVD
Partial estimated background
Negative control samples
Negative control regions
Background
Normalization
Normalized z-scores
z–score
z–score
z–score

Figure 2

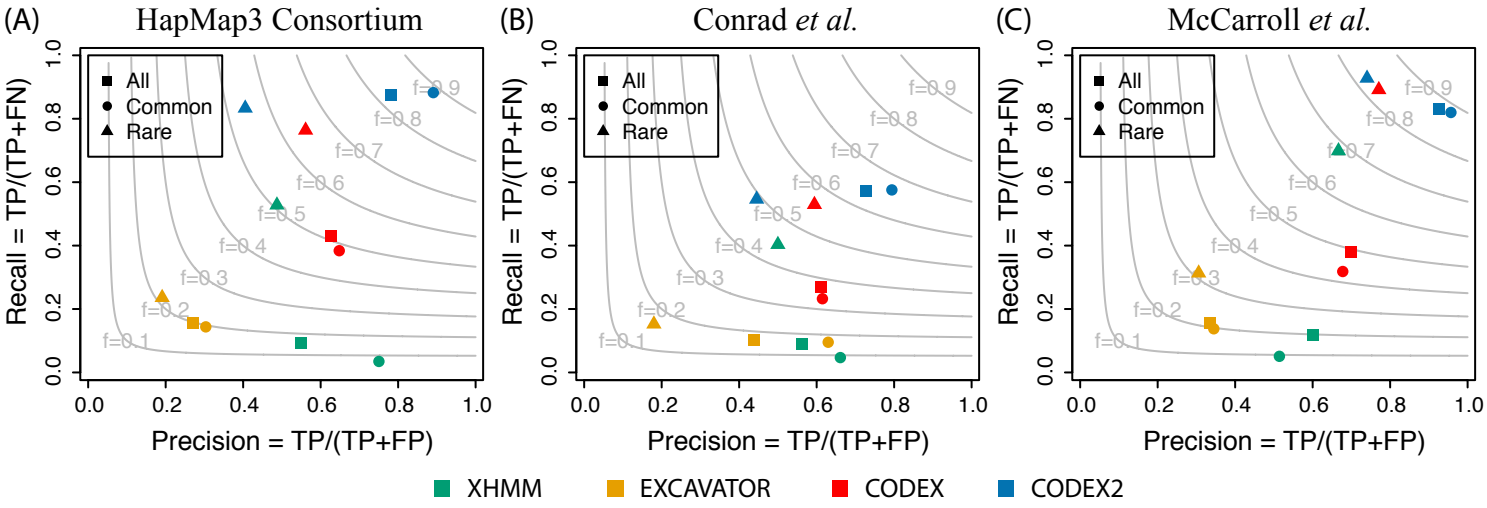

(A)
HapMap3 Consortium
(B)
Conrad et al.
(C)
McCarroll et al.
All
Common
Rare
Recall = TP/(TP+FN)
Precision = TP/(TP+FP)
0.0
0.2
0.4
0.6
0.8
1.0
f=0.1
f=0.2
f=0.3
f=0.4
f=0.5
f=0.6
f=0.7
f=0.8
f=0.9
XHMM
EXCAVATOR
CODEX
CODEX2

Figure 3

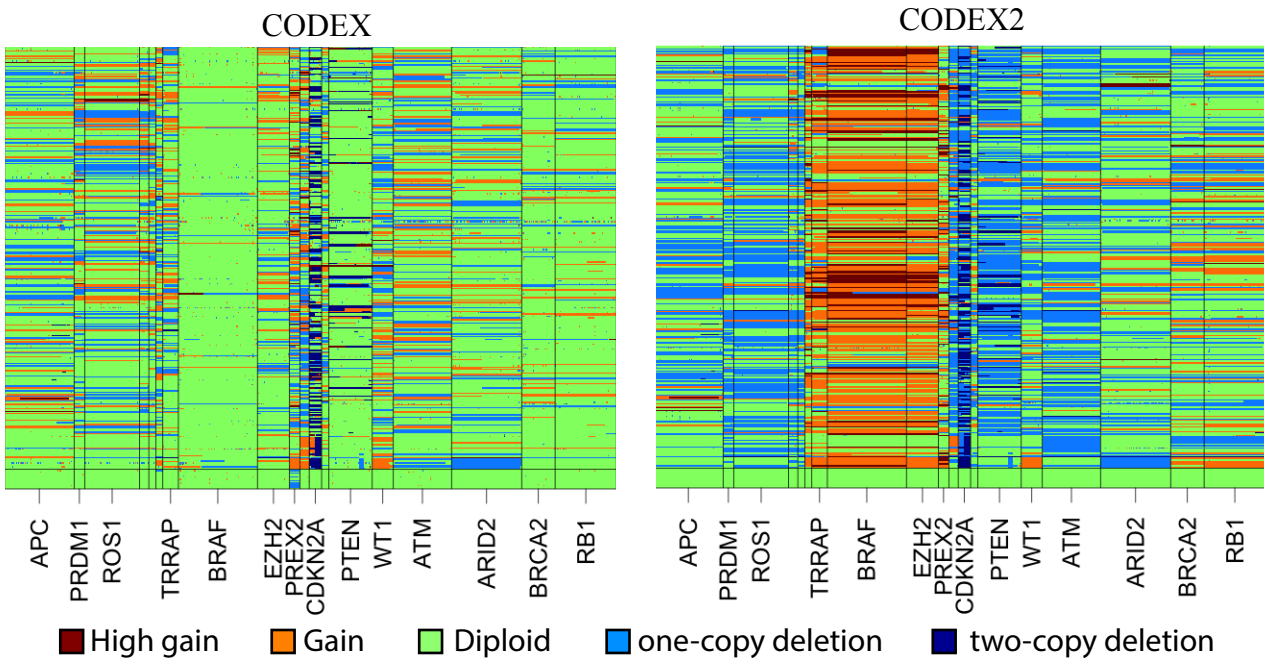

CODEX
CODEX2
APC
PRDM1
ROS1
TRRAP
BRAF
EZH2
PREX2
CDKN2A
PTEN
WT1
ATM
ARID2
BRCA2
RB1
High gain
Gain
Diploid
one-copy deletion
two-copy deletion

Figure 4

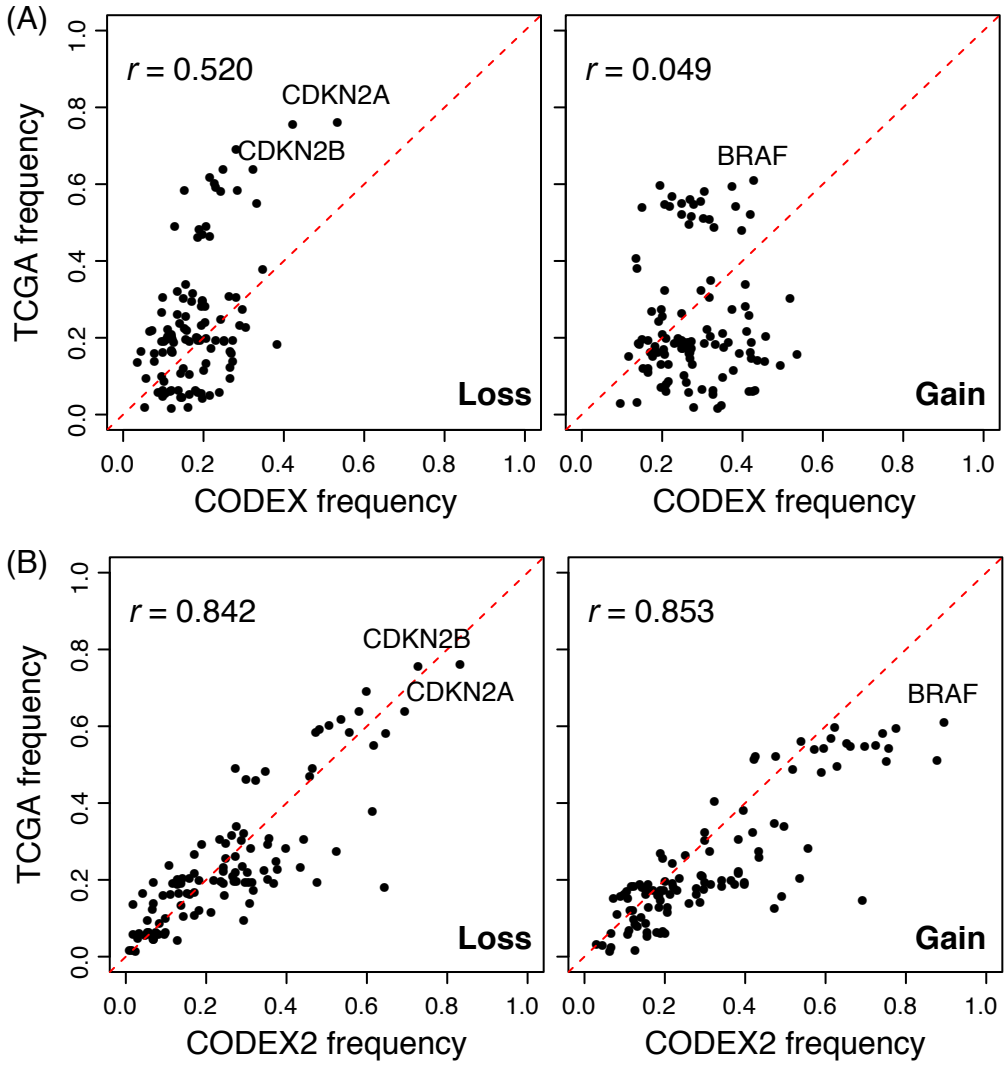

(A)
r = 0.520
CDKN2A
CDKN2B
Loss
TCGA frequency
CODEX frequency
r = 0.049
BRAF
Gain
CODEX frequency
(B)
r = 0.842
CDKN2B
CDKN2A
Loss
TCGA frequency
CODEX2 frequency
r = 0.853
BRAF
Gain
CODEX2 frequency

Figure 5

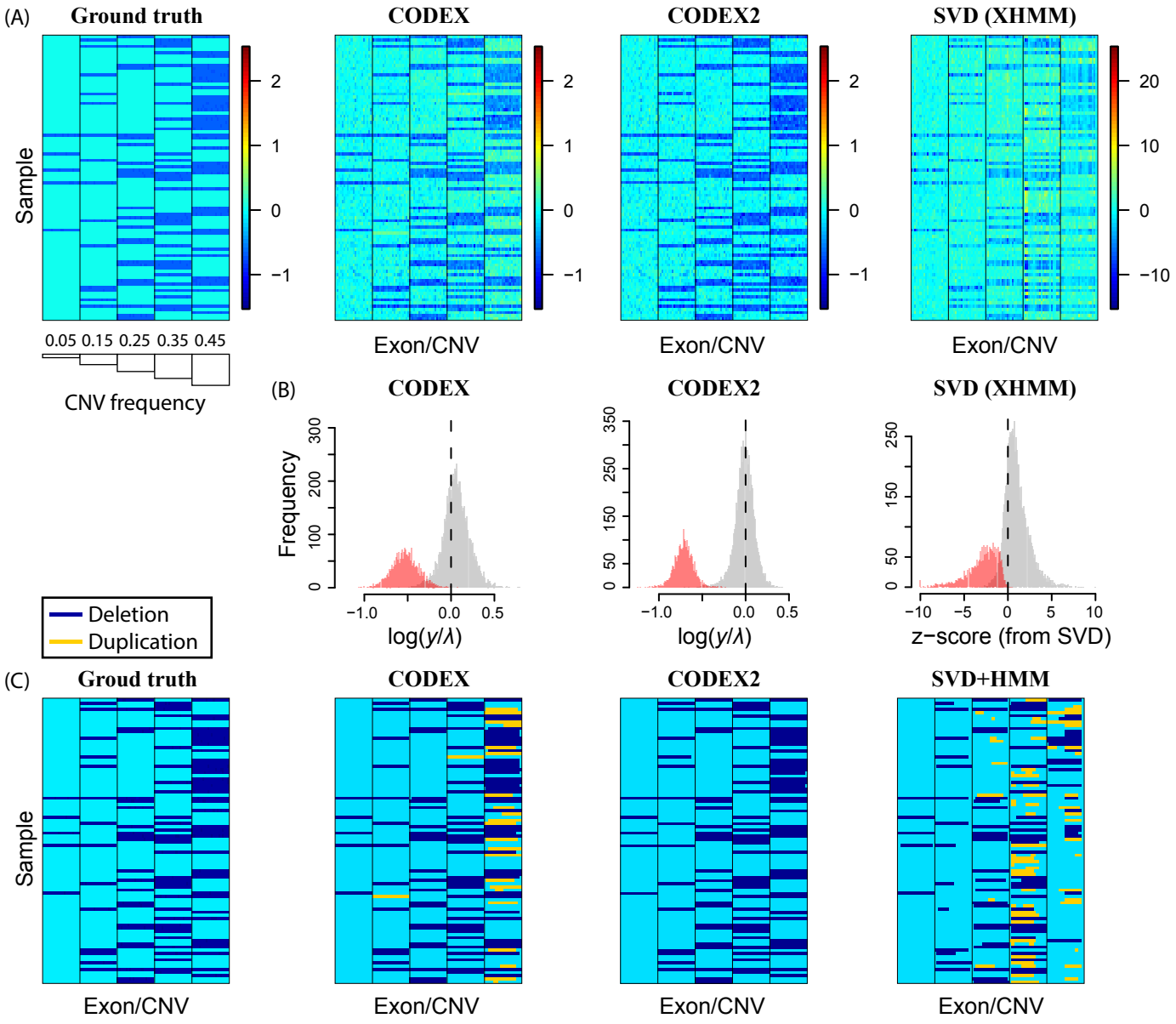

(A)
Ground truth
CODEX
CODEX2
SVD (XHMM)
Sample
Exon/CNV
0.05 0.15 0.25 0.35 0.45
CNV frequency
(B)
Frequency
log(y/λ)
z−score (from SVD)
Deletion
Duplication
(C)
Groud truth
SVD+HMM

Figure 6

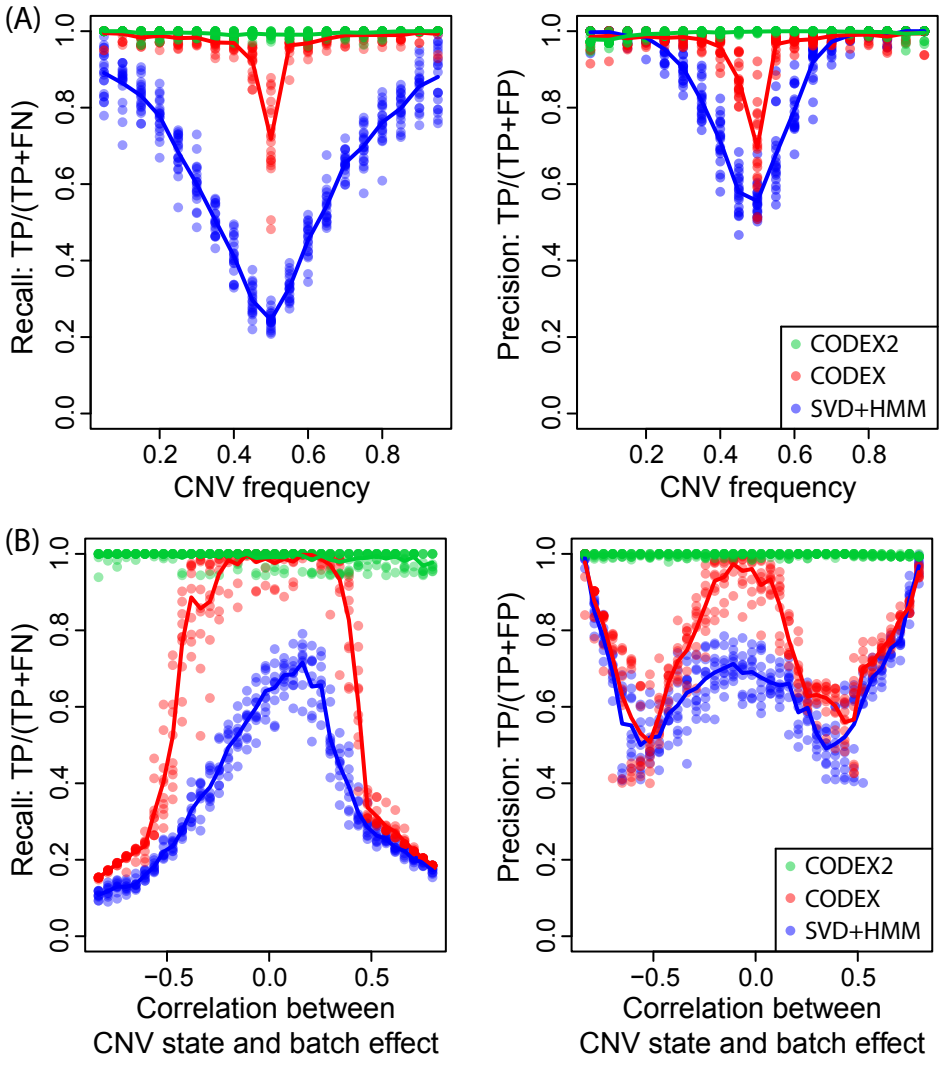

(A)
Recall: TP/(TP+FN)
CNV frequency
Precision: TP/(TP+FP)
CNV frequency
CODEX2
CODEX
SVD+HMM
(B)
Recall: TP/(TP+FN)
Correlation between
CNV state and batch effect
Precision: TP/(TP+FP)
Correlation between
CNV state and batch effect
CODEX2
CODEX
SVD+HMM

Figure 7

Supplement to "CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing"

Yuchao Jiang[1,2,*] Ioannis N. Anastopoulos[3]

Katherine L. Nathanson[3,4] Nancy R. Zhang[5,*]

[1] Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill

[2] Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill

[3] Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine, University of Pennsylvania

[4] Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania

[5] Department of Statistics, The Wharton School, University of Pennsylvania

[*] Corresponding authors: (+1) 919/843-4822; yuchaoj@email.unc.edu; (+1) 215/898-8222; nzh@wharton.upenn.edu.

List of Supplementary Tables

1 Performance of CODEX and CODEX2 with different number of latent factors. 6

List of Supplementary Algorithms

1 Parameter estimation algorithm for CODEX. 3
2 Parameter estimation algorithm for CODEX2 with negative control samples. 4
3 Parameter estimation algorithm for CODEX2 without negative control samples. 5

List of Supplementary Figures

1 Poisson latent factor model for normalizing read depth. 7
2 Biased parameter estimation by CODEX with common CNV signals. 8
3 Model selection and parameter estimation in spike-in studies. 9
4 Assessment of precision and recall rates with different CNV lengths. 10
5 Relationship between CNV state and batch effect. 11

Supplementary Algorithm 1 Parameter estimation algorithm for CODEX.

Initialization:
$\beta^{old} = 1^n$, $g = 0^{n\times K}$, $h = 0^{m\times K}$.
Iteration:

1. For each sample j, fit a smoothing spline $Y_{:j}/N_j\beta^{old}\exp(g \times h_{j:}^T) \sim GC$ to get $f_j(GC)$.
2. For each exon i, update β_i as $\beta_i^{new} = \text{median}\Big([Y/Nf(GC)\exp(g \times h^T)]_{i:}\Big)$.
3. Denote $Z = Nf(GC)\beta^{new}$. Apply SVD to row-centered $\log(Y/Z)$ to obtain the K right singular vectors and use as h^{old}.
   (a) For each i, fit Poisson log-linear regression with $Y_{i:}$ as response, h^{old} as covariates, $\log(Z_{i:})$ as fixed offset to obtain updated estimates as $\{g_{i1}, \cdots, g_{iK}\}$.
   (b) For each j, fit Poisson log-linear regression with $Y_{:j}$ as response, g as covariates, $\log(Z_{:j})$ as fixed offset to obtain updated estimates as $\{h_{j1}^{new}, \cdots, h_{jK}^{new}\}$.
   (c) Center each row of $g \times (h^{new})^T$ and apply SVD to the row-centered matrix to obtain the K right singular vectors to update h^{new}.
   (d) Repeat steps (a) to (c) with $h^{old} = h^{new}$ till convergence to obtain h, g.
4. Repeat steps 1 to 3 with $\beta^{old} = \beta^{new}$ till convergence.

Supplementary Algorithm 2 Parameter estimation algorithm for CODEX2 with negative control samples.

Let J_c be the indices of the control samples.

1. Apply CODEX to the null regions $Y_{:J_c}$and GC using Supplementary Algorithm 1 to get an estimate of $h_{J_c:}, g, \beta$ and $f_{J_c}(GC)$.
2. For case sample $j \notin J_c$, initiate $h_{j1}^{old} = \cdots = h_{jK}^{old} = 0$.
   (a) Fit a smoothing spline $Y_{:j}/N_j\beta \exp(g \times h_{j:}^{old^T}) \sim GC$ to get $f_j(GC)$.
   (b) Denote $Z = Nf(GC)\beta$. Fit Poisson log-linear regression with $Y_{:j}$ as response, g as covariates, $\log(Z_{:j})$ as fixed offset to obtain updated estimates as $\{h_{j1}^{new}, \cdots, h_{jK}^{new}\}$.
   (c) Repeat steps (a) to (b) with $h_{j:}^{old} = h_{j:}^{new}$ till convergence to obtain $h_{j:}$, $f_j(GC)$.

Supplementary Algorithm 3 Parameter estimation algorithm for CODEX2 without negative control samples.

Let I^* be the indices of exons with CNVs with high population frequencies.

1. Apply CODEX to the null regions $Y_{-I^*:}$and GC_{-I^*} using Supplementary Algorithm 1 to get an estimate of $h, g_{-I^*}, \beta_{-I^*}$ and $f(GC_{-I^*})$.
2. Use the estimated non-parametric smooth spline function to estimate the GC content bias in the non-null regions $f(GC_{I^*})$ with the corresponding GC_{I^*}.
3. For each $i^* \in I^*$, adopt an EM algorithm to estimate β_{i^*} and $\{g_{i^*1}, \cdots, g_{i^*K}\}$.
   (a) E-step:

$$\begin{aligned} p_1 &= P\big(Y_{i*j} \mid N_j, f_j(GC_{i^*}), Z_{i^*j} = 1, \hat{\mu}_{i^*}\big)\hat{\pi}_{i^*}, \\ p_0 &= P\big(Y_{i*j} \mid N_j, f_j(GC_{i^*}), Z_{i^*j} = 0, \hat{\mu}_{i^*}\big)(1 - \hat{\pi}_{i^*}), \\ \hat{Z}_{i^*j} &= p_1/(p_1 + p_0), \end{aligned}$$

   where $\hat{\mu}_{i^*}$ and $\hat{\pi}_{i^*}$ are from the previous M-step and $P(Y_{i^*j} \mid \lambda_{i^*j})$ is the probability for Poisson distribution with λ_{i^*j} calculated based on the given parameters.

   (b) M-step:

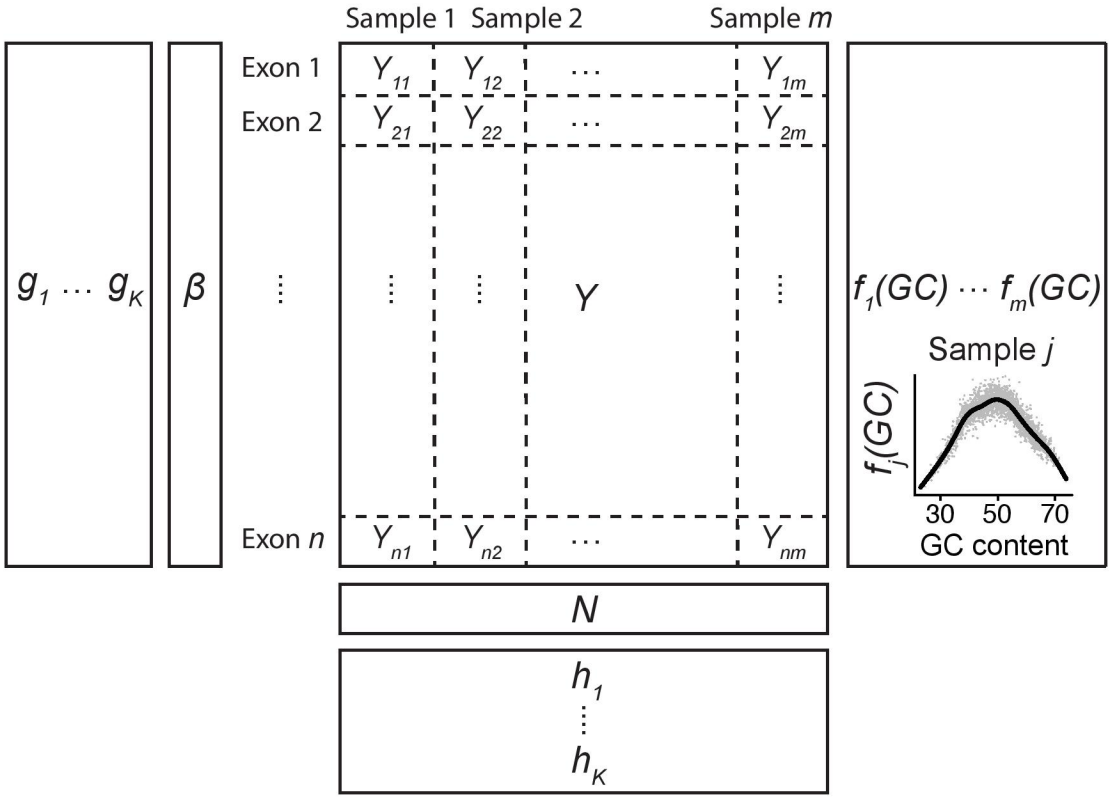
$$\hat{\pi}_{i^*} = \frac{1}{m}\sum_{j=1}^{m} \hat{Z}_{i^*j},$$

   where $\hat{Z}_{i^*j}$ is from the previous E-step. Run Poisson log-linear regression with $Y_{i^*:}$ as response, h and $\hat{Z}_{i^*:}$ as covariates, and $\log(N \times f(GC))_{i^*:}$ as fixed offsets to obtain estimates of $\hat{\beta}_{i^*}$ as intercept, $\{\hat{g}_{i^*1}, ..., \hat{g}_{i^*K}\}$ and $\hat{\mu}_{i^*}$ as coefficients.

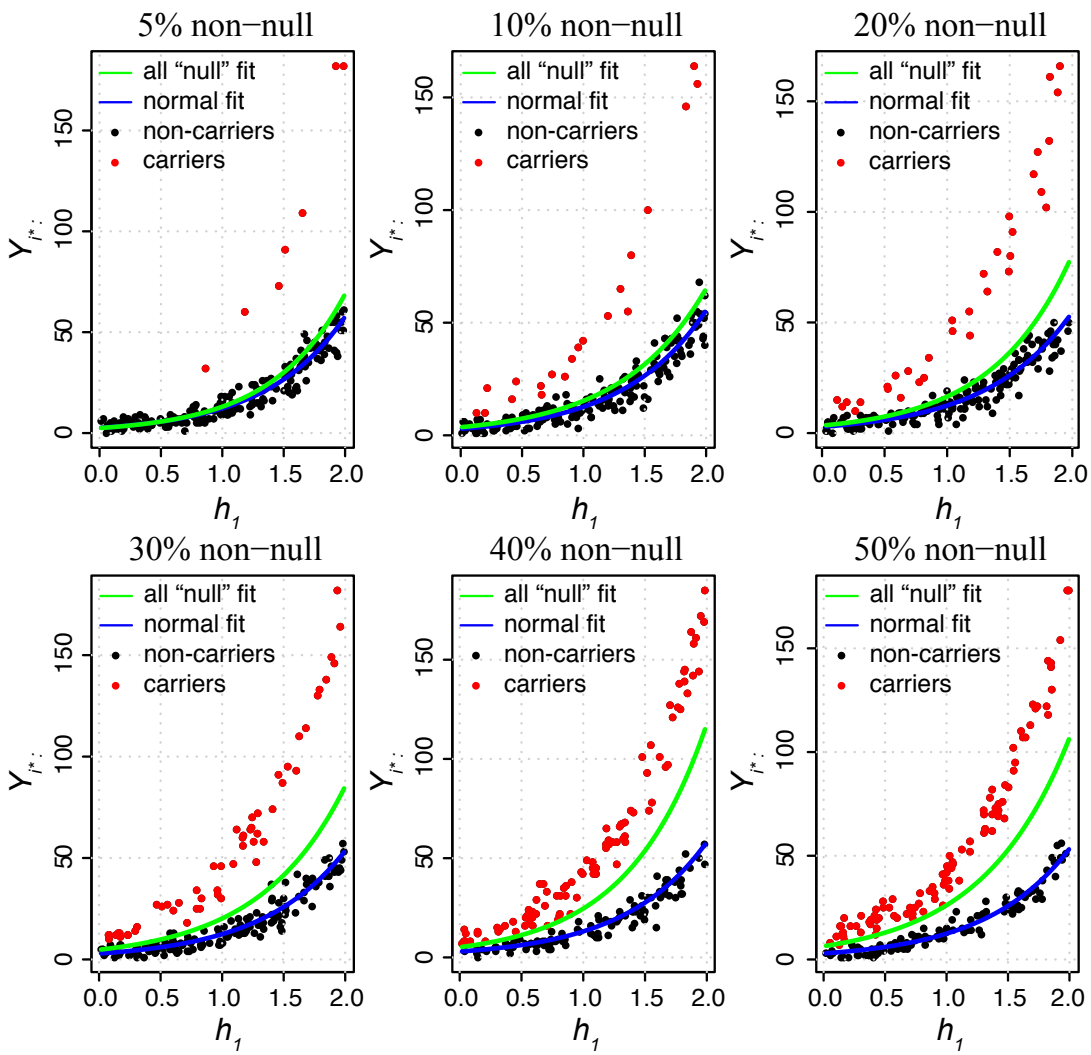
Supplementary Table 1: Performance of CODEX and CODEX2 with different number of latent factors. CODEX and CODEX2 are applied to the melanoma targeted sequencing data set with the number of latent factors K ranging from 0 to 10. Correlations of the profiled losses and gains (r_{loss} and r_{gain} respectively) by CODEX and CODEX2 with those reported by TCGA, as well as the number of *BRAF* gains and *PTEN* losses out of 334 tumor samples are used as measurements. CNV profiles by CODEX2 are consistent since only the negative control samples (16 normal samples) are used to estimate the target-specific bias and artifacts. For CODEX, true CNV signals are attenuated with a large K resulting in less CNV events and lower correlations.

	CODEX2				**CODEX**			
K	r_{loss}	r_{gain}	*BRAF* gains	*PTEN* losses	r_{loss}	r_{gain}	*BRAF* gains	*PTEN* losses
0	0.787	0.859	288	256	0.571	0.112	170	146
1	0.840	0.841	297	230	0.589	-0.039	163	138
2	0.837	0.839	296	228	0.580	-0.003	161	133
3	0.841	0.830	296	227	0.567	0.013	160	117
4	0.837	0.839	292	229	0.581	0.034	158	113
5	0.842	0.848	302	229	0.556	0.046	158	110
6	0.845	0.845	301	229	0.546	0.033	155	108
7	0.842	0.853	298	231	0.520	0.049	142	107
8	0.838	0.859	295	237	0.453	0.027	130	97
9	0.831	0.858	299	234	0.435	0.042	128	95
10	0.835	0.859	297	237	0.442	0.039	125	98

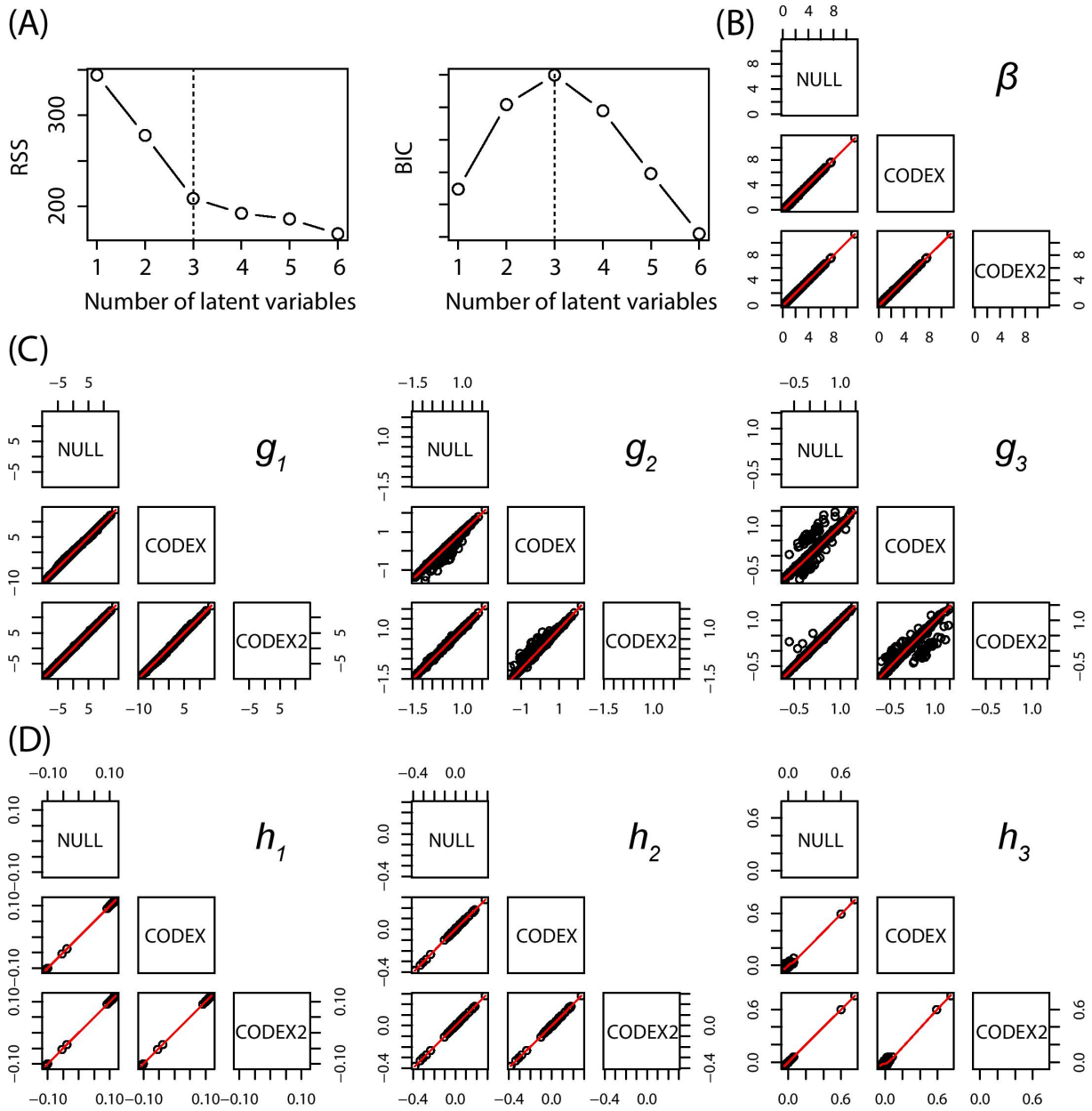
Supplementary Figure 1: Poisson latent factor model to normalize read depth data by whole-exome and targeted sequencing. h^T is shown for illustration purpose.



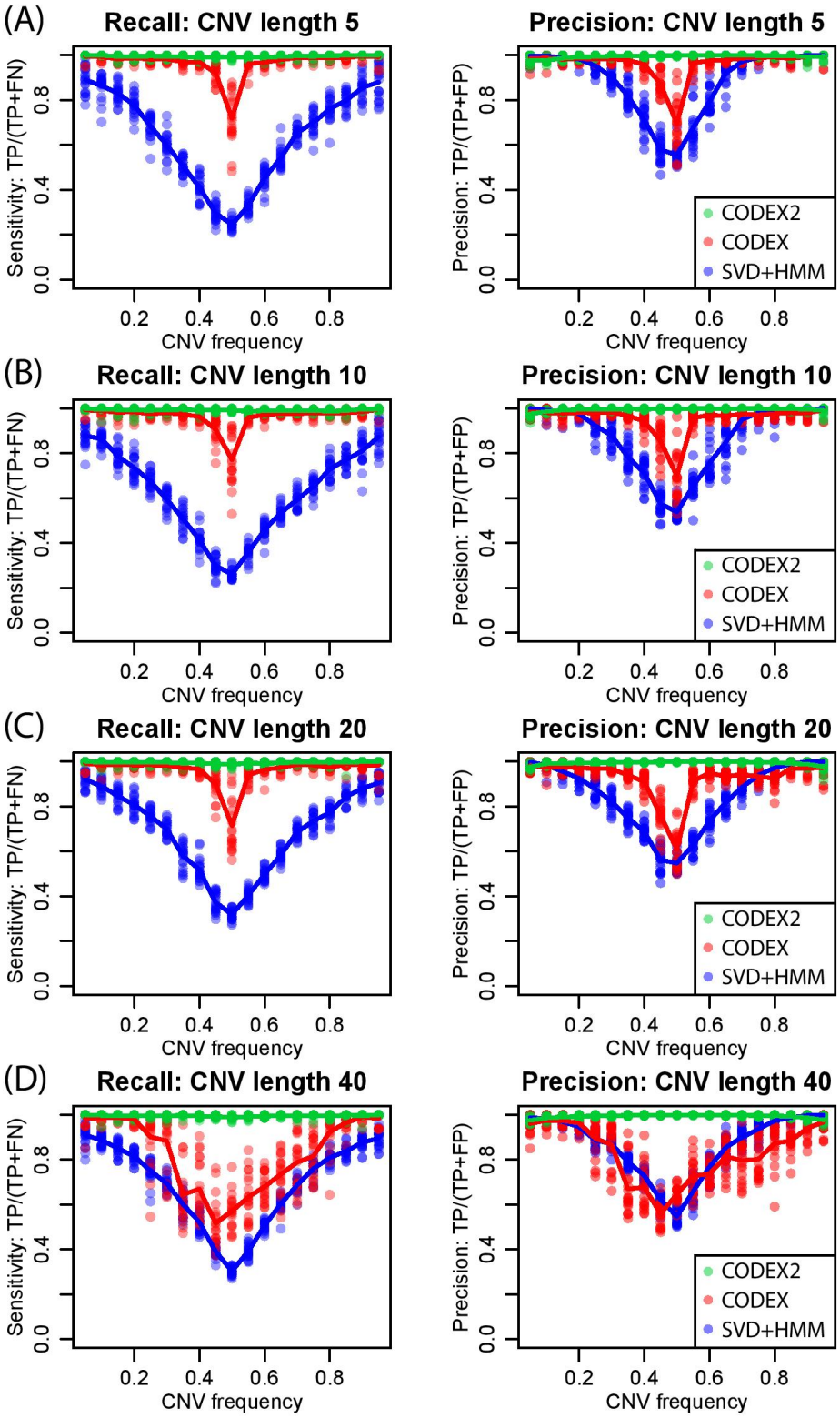
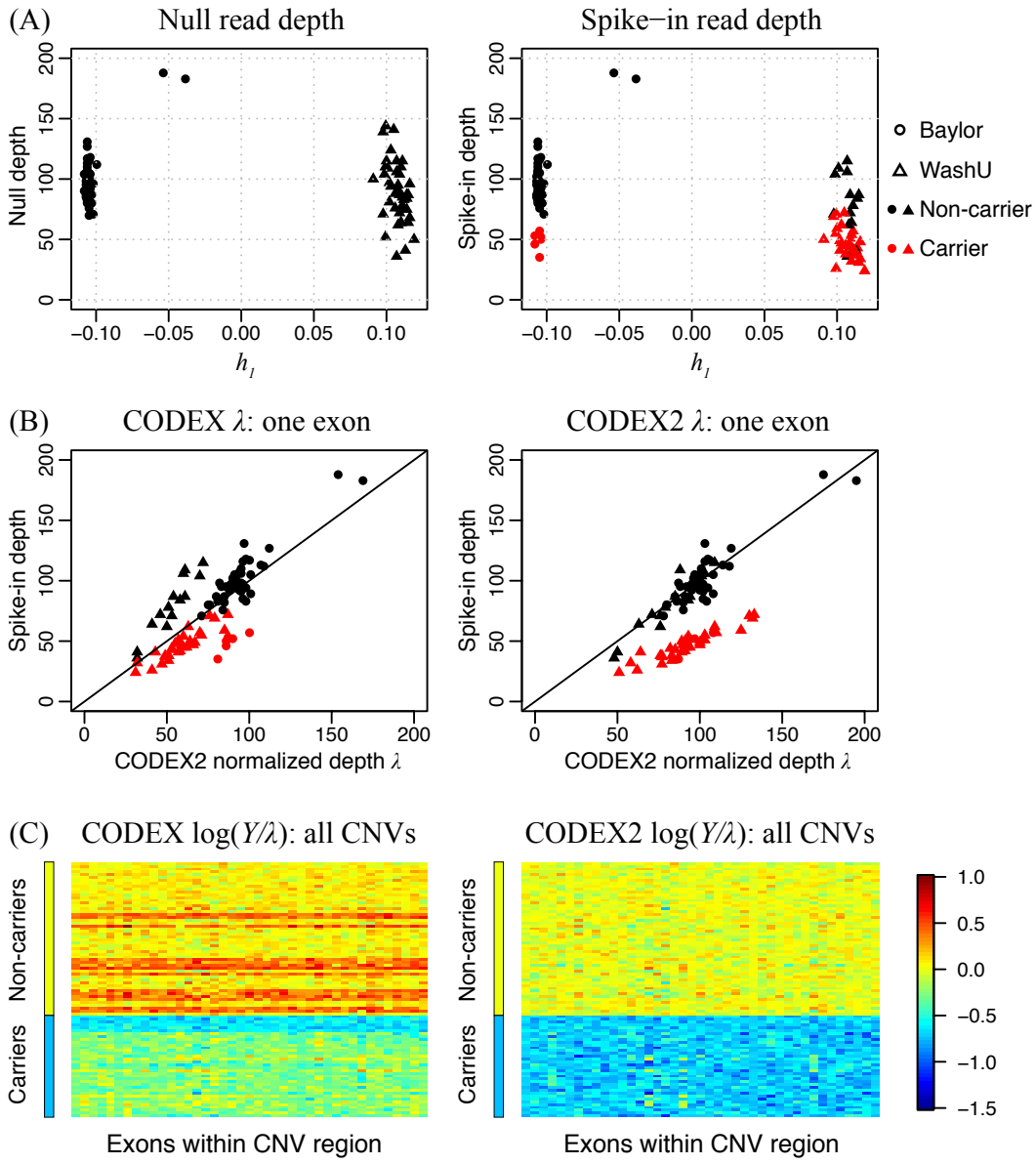
Supplementary Figure 2: Parameter estimation by CODEX when the null data is mixed with alternatives with different frequencies. Here we focus on a specific exon i^* and try to estimate g_{i^*1} from Poisson generalized linear model taking h_1 as known from previous iteration.



Supplementary Figure 3: Model selection and parameter estimation in spike-in studies. (A) Number of latent factors is determined by variance reduction and BIC based on the null read depth after quality control procedures and is kept the same throughout the spike-in studies. (B-D) True underlying parameters and estimated parameters by CODEX and CODEX2. CODEX biasedly estimates the exon-specific latent factor $\{g_1, g_2, g_3\}$ for the exons, where common CNV signals are *in silico* added.



Supplementary Figure 4: Assessment of precision and recall rates with different CNV lengths. CODEX2 has nearly perfect performance compared to CODEX and SVD-based method, which suffer from low recall and precision rates for large and common CNV singals.



Supplementary Figure 5: Relationship between CNV state and batch effect. (A) The null read depth from the 1000 Genomes Project after quality control procedures show distinct batch effect between two centers, where the samples were sequenced. An example spike-in run where the CNVs are mostly added to the samples from one center. The CNV state and the batch effect are highly correlated. (B) An example of Poisson log-linear regression to estimate the exon-specific latent factor for a specific exon in a common CNV region. The vertical axis is Y (in this case the spike-in read depth) while the horizontal axis is the λ (estimated null read depth by CODEX and CODEX2). If there is no CNV, these two quantities should be approximately equal along the diagonal line. CODEX's fitted results separate the carriers and non-carriers at two sides of the diagonal line. CODEX2 adopts the EM algorithm with the non-carriers lying along the diagonal line as expected and the carriers below the diagonal line called as deletions. (C) Heatmap of the normalization results in all non-null exons across all samples. CODEX suffers from low sensitivity for detecting true deletions in carriers and false positives as deletions in non-carriers.