TEST (2016) 25:654–673
DOI 10.1007/s11749-016-0489-z

ORIGINAL PAPER

Limit theorems for empirical Rényi entropy and divergence with applications to molecular diversity analysis

Maciej Pietrzak[1] · Grzegorz A. Rempała[2] ·
Michał Seweryn[3] · Jacek Wesołowski[4]

Received: 9 May 2015 / Accepted: 6 April 2016 / Published online: 20 April 2016

Abstract Quantitative methods for studying biodiversity have been traditionally rooted in the classical theory of finite frequency tables analysis. However, with the help of modern experimental tools, like high-throughput sequencing, we now begin to unlock the outstanding diversity of genomic data in plants and animals reflective of the long evolutionary history of our planet. This molecular data often defies the classical frequency/contingency tables assumptions and seems to require sparse tables with very large number of categories and highly unbalanced cell counts, e.g., following heavy-

Research partially supported by US NIH Grants R01CA-152158 and U01-GM092655 (GAR, MS, MP) as well as US NSF Grant DMS-1318886 (GAR).

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0489-z) contains supplementary material, which is available to authorized users.

✉ Grzegorz A. Rempała
rempala.3@osu.edu

Maciej Pietrzak
pietrzak.20@osu.edu

Michał Seweryn
msewery@math.uni.lodz.pl

Jacek Wesołowski
wesolo@mini.pw.edu.pl

[1] Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA

[2] Division of Biostatistics, College of Public Health and Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA

[3] Department of Mathematics and Computer Science, University of Łódz, Łódz, Poland

[4] Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warsaw, Poland

tailed distributions (for instance, power laws). Motivated by the molecular diversity studies, we propose here a frequency-based framework for biodiversity analysis in the asymptotic regime where the number of categories grows with sample size (an infinite contingency table). Our approach is rooted in information theory and based on the Gaussian limit results for the effective number of species (the Hill numbers) and the empirical Renyi entropy and divergence. We argue that when applied to molecular biodiversity analysis, our methods can properly account for the complicated data frequency patterns on one hand and the practical sample size limitations on the other. We illustrate this principle with two specific RNA sequencing examples: a comparative study of T-cell receptor populations and a validation of some preselected molecular hepatocellular carcinoma (HCC) markers.

Keywords Hill number · Central limit theorem · Next-generation sequencing · Triangular arrays · T-cell receptors

Mathematics Subject Classification 60F05 · 60G42 · 94A17

1 Introduction

Developing effective methods for quantifying and comparing empirical diversity of various biological populations is one of the fundamental problems of modern life sciences, as it has direct impact on our understanding of the basic operating principles of our planet's ecosystem and its evolution (cf., eg., Berkov et al. 2014). In the course of its 3.5 billion years of evolutionary history, nature has developed an outstanding bio- and molecular diversity among the Earth's species of plants and animals. Indeed, it is estimated that there are currently about 8.7 million eukaryotic species on earth, both marine and terrestrial, 88 % of which are still waiting to be described (Mora et al. 2011). The diversity at the molecular level is perhaps even more spectacular, as it occurs at different levels of biological organization: within one individual (e.g., through RNA, DNA, proteins, and metabolites), between individuals of the same and related species, within and between species and ecosystems, as well as throughout evolution (see, e.g., Campbell 2003). For instance, the number of different molecular types of human T cells is estimated at 10^{18} (Cea 2005) which is only slightly less than the currently estimated number of stellar objects in the known universe (the latter believed to be of the order 10^{21}).

Whereas the power of modern computing has allowed us to make steady progress toward building even more robust empirical measures of biodiversity based on a variety of considerations (see, e.g., Presley et al. 2014), the most relevant to our discussion here are the measures borrowed from the field of information theory. They include among others the *Hill number* (or the effective number of species) and the related concept of the *Renyi entropy* (see, e.g., the recent review Chiu et al. (2014) and references therein). Although originally proposed for quantifying ecological diversity in the macroscale ecosystems (Chao et al. 2010), the use of the empirical Renyi entropy as a descriptor of diversity was also adopted for molecular populations in de Andrade and Wang (2011). Since then, the Renyi-type measures have been applied to problems of molecular populations ranging from analyzing regulatory variants and testing genome-wide associations (Sun and Hu 2013; Sadee et al. 2014) to comparing different T-cell populations (Cebula et al. 2013; Rempala and Seweryn 2013). Despite their growing usage in biodiversity studies of both macro- and molecular-level populations, it appears that some important statistical properties of the Renyi-type measures have not been yet sufficiently understood, especially in the context of frequency-based analysis and large sample behavior.

Currently, the standard methods of obtaining molecular-level data on the *transcriptome* (RNA) abundance rely on the so-called next-generation sequencing (NGS) technology and especially on the high-throughput RNA sequencing or RNA-seq (Wang et al. 2009). However, the molecular count data from NGS often elude standard statistical analysis due to the fact that exhaustive sampling of the DNA and RNA fragments for the purpose of sequence reconstruction is not feasible and the sequencing errors increase with sampling intensity or *sequencing depth* (O'Rawe et al. 2015). It has been therefore generally conceded (Oh et al. 2014) that the standard, fixed-dimension, non-parametric frequency/contingency table analysis (see, e.g., Agresti 2002) does not readily apply to the NGS data and that a different, *infinite-size* contingency table framework, more reflective of the current sequencing technology, appears necessary. Due to the nature of the NGS methods, such framework should be based on the large sample (high-throughput) considerations but, at the same time, should also account for the increase in the number of sequencing errors with increasing sample size as well as for the undersampling bias.

Motivated by the questions on comparing biodiversity in molecular data (especially, arriving from the NGS experiments) in the current paper, we establish some large sample results for the empirical Renyi entropy and divergence to bridge the gap between the current heuristic approaches and a more formal statistical theory of large samples. To this end, we derive herein several central limit theorems (CLTs) which yield approximate confidence bounds for the (Renyi) entropy-based measures of diversity and similarity in the setting of an infinite contingency table. Our CLT results complement both the law of large number theorems in Rempala and Seweryn (2013) and the CLT for the plugin estimates of the Shannon entropy Zhang and Zhang (2012) and the Kullback–Leibler divergence estimates (Paninski 2003; Zhang and Grabchak 2014). Since in the NGS experiments, one typically expects to undersample the transcriptome, we focus here on the Renyi entropy exponent (which is denoted below by α) less than one so as to upweight the contributions of the lower counts, and our CLT results are restricted to this case. The extensions to arbitrary exponents are straightforward, but not considered here. To provide examples of the types of applications motivating the mathematical results, we analyze two real biological datasets from two different types of NGS experiments. In the first experiment, described in the study Cebula et al. (2013), one compares multiple T-cell receptors populations taken from mice before and after treatment with antibiotics. The goal of the second experiment is the elucidation of differences in gene expression profiles between cancer and control tissues in individuals with hepatocellular carcinoma, as described in Chan et al. (2014). In both presented examples, the NGS datasets are analyzed and de-noised by applying a multi-stage process developed on the basis of our theoretical results.

As already indicated above, the problem of empirically estimating entropy and divergence has been extensively studied in the statistical and machine learning literature over the past several decades, both in the context of discrete and continuous distributions. See, for instance, the monograph by Pardo (2005) or the review in Krishnamurthy et al. (2014) for more details. In the general case of Renyi's entropy and closely related Tsallis' entropy of a fixed continuous distribution f in $\mathbb{R}^m$, a class of consistent estimators was proposed in Leonenko et al. (2008) based on the k-th nearest-neighbor distances computed from the appropriate random samples of size n from f. The idea was later also extended to the Renyi entropy functionals in Källberg et al. (2012) and it appears that similar results could be expected to hold in the discrete case as well. The main difference between these types of results and what is considered here is that in our setting, the discrete density function f is allowed to change as the sample size n increases and that we only analyze the basic empirical frequency (the so-called plug-in) estimates.

The paper is organized as follows. In the next section (Sect. 2) we outline the relevant mathematical concepts along with the necessary notation. In Sect. 3 we state the main theoretical results of the paper, namely the CLTs for the Hill number (or the Tsallis entropy) and the Renyi entropy and divergence in the asymptotic regime when the diversity of the population (i.e., the number of different types) grows with the sample size. The results for the simpler case (Theorems 1 and 2), when Renyi entropy statistics admit linear approximations, are established via the intermediate CLT results for the corresponding power sums, which are closely related to the CLTs for Hill's numbers and Tsallis' entropies. These results are also included as parts of formulations of Theorems 1 and 2. In case of the uniform distribution for the Renyi entropy as well as the equal-marginal bivariate distribution for the Renyi divergence, the power sum CLTs are no longer valid (there is no linear approximation available) and other methods are required to establish weak convergence to Gaussian variates under slightly more stringent conditions. These results are presented as Theorems 3 and 4 in Sect. 3. As it turns out, the key ingredient needed to establish Theorems 3 and 4 is the CLT result for two Pearson-type Chi-square statistics in an infinite contingency table. This latter result is of interest in itself and is presented as Lemma 2 in Sect. 3. In the following Sect. 4, we provide some simulation-based examples of the asymptotic behavior of estimates from Sect. 3 in the case (relevant to our applications) of power law distributions under various sampling scenarios. These examples illustrate in particular how the CLTs of Sect. 3 may hold or not, depending on the relations between the dimensions of the relevant contingency tables and the empirical sample sizes. In the second part of Sect. 4, we also discuss in detail two biological examples of NGS data analysis and show how the results of Sect. 3 may be used to analyze the biodiversity of T-cell receptors and to profile the multiple sets of transcriptomes. The final Sect. 5 offers a summary and brief conclusions. The proofs of all more complicated results are provided in the appendix along with some auxiliary technical lemmas.

2 Power sums, entropy and divergence

Consider a triangular array of bivariate row-wise independent random variables $Z_{n,k}$ for $k = 1, \ldots, n$ which in each row are equidistributed with the random variable $Z_n = (X_n, Y_n)$, such that $P(X_n = i, Y_n = j) = p_{ij}^{(n)}$ for $i, j = 1, \ldots, m_n$. Below, we suppress the index n when possible, writing, m, Z_k, Z, p_{ij}, etc. for simplicity.

Let $\alpha > 0$ and for any probability distribution $\boldsymbol{p} = (p_i)_{i=1}^m$ define

$$\mathcal{S}_\alpha(\boldsymbol{p}) = \sum_{i=1}^{m} p_i^\alpha. \tag{2.1}$$

Similarly, for any pair of distributions $\boldsymbol{p} = (p_i)_{i=1}^m$ and $\boldsymbol{q} = (q_i)_{i=1}^m$, define

$$\mathcal{S}_\alpha(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{m} p_i^\alpha q_i^{1-\alpha}. \tag{2.2}$$

(Note that $\mathcal{S}_1 \equiv 1$). The well-known special case of the above is $\alpha = 1/2$, which results in a symmetric index $\mathcal{S}_{1/2}(\boldsymbol{p}, \boldsymbol{q}) = \mathcal{S}_{1/2}(\boldsymbol{q}, \boldsymbol{p})$ often referred to as the Bhattacharyya coefficient (see, e.g., Nielsen and Boltz 2011).

Recall (Renyi 1961) that for a given distribution $\boldsymbol{p}$, its Renyi entropy $\mathcal{H}_\alpha$ is defined as

$$\mathcal{H}_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \left(\sum p_i^\alpha \right) = \frac{1}{1-\alpha} \log \mathcal{S}_\alpha(\boldsymbol{p}),$$

and that for a pair of distributions $(\boldsymbol{p}, \boldsymbol{q})$, their Renyi divergence $\mathcal{D}_\alpha$ is defined as

$$\mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\alpha - 1} \log \mathcal{S}_\alpha(\boldsymbol{p}, \boldsymbol{q}).$$

Note that the sign change in the normalizing constant is needed to ensure non-negativity of $\mathcal{H}_\alpha$ and $\mathcal{D}_\alpha$. The special case of $\mathcal{D}_\alpha$ with $\alpha = 1/2$ is referred to as the Bhattacharyya distance and may be expressed in terms of the Mahalanobis distance (see, e.g., Nielsen and Boltz 2011), whereas the linear approximation of $\mathcal{H}_\alpha(\boldsymbol{p})$ given by

$$\mathcal{T}_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha}(\mathcal{S}_\alpha(\boldsymbol{p}) - 1) \tag{2.3}$$

is sometimes referred to as the Tsallis entropy and has important applications in the field of statistical mechanics (Tsallis 1988). Note that for our current purposes, we will only consider the quantities $\mathcal{D}_\alpha$, $\mathcal{H}_\alpha$, and $\mathcal{T}_\alpha$ for α satisfying $0 < \alpha < 1$.

In what follows, the summation symbol without subscripts ($\sum$) will indicate summation with respect to the index i ($i = 1, \ldots, m$), whereas $\boldsymbol{p} = (p_i)_{i=1}^m$ and $\boldsymbol{q} = (q_i)_{i=1}^m$ will (typically) denote the marginal distributions of the bivariate variable

$Z = (X, Y)$ whose distribution is denoted by $(p_{ij})_{i,j=1}^m$. Additionally, the uniform distribution on m points will be denoted by $\boldsymbol{u}$. An important relation between the Renyi entropy and the Renyi divergence is

$$\mathcal{H}_\alpha(\boldsymbol{p}) = \log m - \mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{u}). \tag{2.4}$$

We note also the following monotonicity property of $\mathcal{D}_\alpha$ and $\mathcal{H}_\alpha$ with respect to the index α.

Lemma 1 *For* $0 < \alpha < \beta < 1$, *we have* $\mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) \le \mathcal{D}_\beta(\boldsymbol{p}, \boldsymbol{q})$ *and, thus, in view of* (2.4)*, also* $\mathcal{H}_\alpha(\boldsymbol{p}) \ge \mathcal{H}_\beta(\boldsymbol{p})$.

Proof Note that for $x \ge 0$, the function $x \to x^{\frac{\alpha-1}{\beta-1}}$ is strictly convex for $0 < \alpha < \beta < 1$. Therefore, by Jensen's inequality,

$$\begin{aligned}\mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) &= \frac{1}{\alpha-1} \log \sum p_i^\alpha q_i^{1-\alpha} = \frac{1}{\alpha-1} \log \sum p_i \left(\frac{q_i}{p_i}\right)^{(1-\beta)\frac{\alpha-1}{\beta-1}} \\ &\le \frac{1}{\beta-1} \log \sum p_i \left(\frac{q_i}{p_i}\right)^{(1-\beta)} = \mathcal{D}_\beta(\boldsymbol{p}, \boldsymbol{q}).\end{aligned}$$

Example 1 (Hill's Number) For a given $0 < \alpha < 1$, the measure of diversity of a distribution $\boldsymbol{p}$ also known as the *effective number of classes* may be defined as (see, e.g., Jost 2007; Chao et al. 2012; Rempala and Seweryn 2013) $ENC_\alpha(\boldsymbol{p}) = \exp(\mathcal{H}_\alpha(\boldsymbol{p})) = \mathcal{S}_\alpha(\boldsymbol{p})^{1/(1-\alpha)}$. It follows then from Lemma 1 that for any $0 < \alpha < \beta < 1$, we have $ENC_\alpha(\boldsymbol{p}) \ge ENC_\beta(\boldsymbol{p})$. (As it turns out, this inequality may be in fact extended to arbitrary positive $\alpha < \beta$).

2.1 Low diversity condition and projection variables

The notion of an infinite-dimension contingency table brought up in the introduction may be now formally introduced simply as a requirement that for n-size sample from $(p_{ij})_{i,j=1}^m$ we have $m \to \infty$ as $n \to \infty$. Throughout the paper, let $a \wedge b$ denote $\min(a, b)$ for any real a, b and let $a_n \sim b_n$ (resp. $a_n \sim O(b_n)$) denote $a_n/b_n \to 1$ (resp. $A < \limsup_n a_n/b_n < B$ for some finite A, B) as $n \to \infty$ for any real sequences a_n, b_n. Throughout the paper, we consider only the *low diversity* (LD) schemes in which the marginals $\boldsymbol{p}, \boldsymbol{q}$, of Z satisfy the following *LD condition.*

$$(np_*)^{-1} = o(n^{-\tau}) \quad \text{for some} \quad \tau > 0, \tag{2.5}$$

where $p_* = \min_i(p_i) \wedge \min_i(q_i)$. Note that since $p_* \le 1/m$. (2.5) implies in particular $m/n = o(n^{-\tau})$. As it turns out, for many distributions $\boldsymbol{p}$, the two conditions are in fact equivalent, as seen in the following.

Example 2 (Power Law Model) Let $\boldsymbol{p} = \boldsymbol{q}$ and assume that $p_i = H^{-1}(\beta, m)/(i^\beta l(i))$, $(i = 1, \ldots, m)$ where $\beta > 0$, $l(x)$ is a non-decreasing slowly varying function (see, e.g., Soulier 2009, chapter 1), and $H^{-1}(\beta, m) = 1/\sum_{i=1}^{m}(i^{\beta}l(i))^{-1}$ is the normalizing constant. Note that if $0 < \beta < 1$, then $H^{-1}(\beta, m) \sim (1-\beta)l(m)/m^{1-\beta}$ and (2.5) is implied by $m/n = o(n^{-\tau})$, since

$$(n \min_i p_i)^{-1} \sim (1-\beta)^{-1}\frac{m^{\beta}l(m)}{nm^{\beta-1}l(m)} = (1-\beta)^{-1}\frac{m}{n}.$$

For any $0 < \alpha < 1$ and a given pair (m, n), let us define two random variables which will play an important role in the following section. Let $W_n^{(\alpha)}$ be defined as

$$P\left(W_n^{(\alpha)} = \alpha p_i^{\alpha-1}\right) = p_i \tag{2.6}$$

for $i = 1, \ldots, m$. Similarly, define also $V_n^{(\alpha)}$ as

$$P\left(V_n^{(\alpha)} = \alpha\left(\frac{q_i}{p_i}\right)^{1-\alpha} + (1-\alpha)\left(\frac{p_j}{q_j}\right)^{\alpha}\right) = p_{ij} \tag{2.7}$$

for $i, j = 1, \ldots, m$. In the following, for the reasons discussed below, we refer to (2.6) and (2.7) as the *projection variables* or simply *projections*.

Remark 1 Note that

$$EW_n^{(\alpha)} = \alpha\mathcal{S}_{\alpha}(\boldsymbol{p})$$

and $VarW_n^{(\alpha)} = 0$ iff $p_i = 1/m$ for all i, that is, $\boldsymbol{p} = (p_i) = \boldsymbol{u}$ is a uniform distribution on m support points (this case is often referred to as a maximal diversity model or a pure noise model). Similarly,

$$EV_n^{(\alpha)} = \mathcal{S}_{\alpha}(\boldsymbol{p}, \boldsymbol{q})$$

and it is also easy to see that $VarV_n^{(\alpha)} = 0$ iff $p_i = q_i$ for all i, that is, $\boldsymbol{p} = \boldsymbol{q}$.

As it turns out, both cases $\boldsymbol{p} = \boldsymbol{u}$ and $\boldsymbol{p} = \boldsymbol{q}$ require special consideration in the asymptotic analysis of $\mathcal{H}_{\alpha}$ and $\mathcal{D}_{\alpha}$. In view of the remark above, they are referred to in what follows as the cases of “degenerate” (zero variance) projections.

Example 3 (Noise-and-signal and pure noise models) A distribution concentrated on $m+1$ support points, such that $p_0 > 0$ and $p_i = (1-p_0)/m$ for $1 \le i \le m$, may be considered as a simple model of signal contamination. Note that in this case, we have $P(W_n^{(\alpha)} = \alpha p_0^{\alpha-1}) = p_0$, $P(W_n^{(\alpha)} = \alpha m^{1-\alpha}(1-p_0)^{\alpha-1}) = 1-p_0$ and

$$VarW_n^{(\alpha)} = \alpha^2\left(m^{1-\alpha}(1-p_0)^{\alpha}\left(\frac{p_0}{1-p_0}\right)^{1/2} - p_0^{\alpha}\left(\frac{1-p_0}{p_0}\right)^{1/2}\right)^2.$$

For the pure noise model $p_0 = 0$, in which case the support reduces to m points, the above formula is not valid. However, as already pointed out before, in this case we may show directly that $Var W_n^{(\alpha)} = 0$.

3 Limit theorems

Let $N(0, 1)$ denote the standard Gaussian random variable and $\Rightarrow$ denote the usual weak convergence in the space of probability distributions. Define also the plug-in n-sample estimates of $\boldsymbol{p}$ and $\boldsymbol{q}$ as, respectively, $\hat{\boldsymbol{p}} = (\hat{p}_i)_{i=1}^m$, where $\hat{p}_i = \sum_{k=1}^n I(X_k = i)/n$ and $\hat{\boldsymbol{q}} = (\hat{q}_i)_{i=1}^m$, where $\hat{q}_i = \sum_{k=1}^n I(Y_k = i)/n$. Here and elsewhere in the paper, $I(\cdot)$ denotes the indicator function. As it turns out, two distinct sets of CLTs may be derived depending on whether the variables $W_n^{(\alpha)}$ and $V_n^{(\alpha)}$ are degenerate (that is, their respective variances vanish) or not. For the non-degenerate case, the appropriate CLTs may be established by expanding on the usual projection and Taylor's expansion arguments (see, e.g., Shao 2003, chapter 1) as well as some elementary bounds on binomial moments (Knoblauch 2008). This is a simpler case to consider and we discuss it first.

3.1 CLTs for non-degenerate projections

The first two CLT results for the empirical (plug-in) Renyi entropy and divergence and their corresponding power sums are provided in Theorems 1 and 2 below. Their respective hypotheses (iii) may be viewed as complementing the analogous results established for the Shannon entropy and the Kullback–Leibler divergence (Paninski 2003; Zhang and Zhang 2012; Zhang and Grabchak 2014). Note also that $\mathcal{S}_\alpha = (ENC_\alpha)^{1-\alpha}$ where the Hill number ENC_α is defined in Example 1. The proofs are deferred to the appendix.

Recall that for any square integrable random variable X, such that $EX \neq 0$, we define its coefficient of variation as $\mathcal{CV}(X) = (Var X)^{1/2}|EX|^{-1}$.

Theorem 1 (Renyi Entropy CLT) *Let $W_n^{(\alpha)}$ be a sequence of random variables defined by (2.6) such that $\mathcal{CV}(W_n^{(\alpha)}) > 0$ and let*

$$\sum p_i^{\alpha-1}(n Var W_n^{(\alpha)})^{-1/2} \to 0 \quad \textit{for } m, n \to \infty. \tag{3.1}$$

Then, under the LD condition (2.5), as $m, n \to \infty$,

- *(i)* $\mathcal{S}_\alpha(\hat{\boldsymbol{p}})/\mathcal{S}_\alpha(\boldsymbol{p}) \to 1$ *in probability,*
- *(ii)* $\sqrt{n}(\mathcal{S}_\alpha(\hat{\boldsymbol{p}}) - \mathcal{S}_\alpha(\boldsymbol{p}))/(Var W_n^{(\alpha)})^{1/2} \Rightarrow N(0, 1)$,
- *(iii)* $\sqrt{n}\,(1/\alpha - 1)(\mathcal{H}_\alpha(\hat{\boldsymbol{p}}) - \mathcal{H}_\alpha(\boldsymbol{p}))/\mathcal{CV}(W_n^{(\alpha)}) \Rightarrow N(0, 1)$.

Remark 2 Note that the first two assertions of the theorem may be equivalently stated in terms of the convergence of the Tsallis plug-in entropy defined by (2.3).

Remark 3 Note that the condition (3.1) is typically stronger than (2.5). Indeed, taking $\alpha > 1/2$ and the power law model from Example 2 with $0 < \beta < 1$, we obtain

$\sum p_i^\alpha \sim (1-\beta)^\alpha m^{1-\alpha}/(1-\alpha\beta)$ and $\sum p_i^{2\alpha-1} \sim (1-\beta)^{2\alpha-1} m^{2-2\alpha}/(1-2\alpha\beta+\beta)$. Consequently, for some constant $C > 1$

$$\frac{C\sum p_i^{\alpha-1}}{\sqrt{n\left(\sum p_i^{2\alpha-1}-(\sum p_i^\alpha)^2\right)}} \geq \frac{m}{\sqrt{n}} \frac{(\max_i p_i)^{\alpha-1}}{m^{1-\alpha}} \geq \frac{m}{\sqrt{n}}$$

for large m, n, and (3.1) implies (2.5) with $\tau = 1/2$. Similarly, (possibly for different $C > 1$),

$$\frac{\sum p_i^{\alpha-1}}{\sqrt{n\left(\sum p_i^{2\alpha-1}-(\sum p_i^\alpha)^2\right)}} \leq \frac{Cm}{\sqrt{n}} \frac{(\min_i p_i)^{\alpha-1}}{m^{1-\alpha}} \leq C\frac{m}{\sqrt{n}}$$

and therefore in this case, (3.1) is seen to be equivalent to (2.5) with $\tau = 1/2$.

Remark 4 (Plug-in bias) Note that, in view of Jensen's inequality applied to the strictly concave function $x \to x^\alpha$ for $x > 0$ and $0 < \alpha < 1$, we have $E\mathcal{S}_\alpha(\hat{\boldsymbol{p}})/\mathcal{S}_\alpha(\boldsymbol{p}) \leq 1$. This and the assertion (i) above imply together that under the assumptions of Theorem 1, the *relative bias* of $\mathcal{S}_\alpha(\hat{\boldsymbol{p}})$ satisfies $E\mathcal{S}_\alpha(\hat{\boldsymbol{p}})/\mathcal{S}_\alpha(\boldsymbol{p}) - 1 \to 0$ as $n, m \to \infty$. The standard inequality $\log x \leq x - 1$ valid for $x > 0$ implies then that the bias of the plug-in entropy estimate satisfies

$$E\mathcal{H}_\alpha(\hat{\boldsymbol{p}}) - \mathcal{H}_\alpha(\boldsymbol{p}) \to 0 \quad \text{as} \quad n, m \to \infty. \tag{3.2}$$

Unfortunately, as may be seen from the proof of Theorem 1 in the appendix, a more careful analysis of the tail events for the plug-in estimate than the one currently performed is needed to establish the actual convergence rate in (3.2).

Turning now to our second result, note that the relation (2.4) suggests that CLT of Theorem 1 could be also extended to the Renyi divergence. The proof is again based on the Taylor expansion method where now the projection variable (2.6) is replaced by (2.7).

Theorem 2 (Renyi Divergence CLT) *Let $V_n^{(\alpha)}$ be a sequence of random variables defined by* (2.7) *such that $\mathcal{CV}(V_n^{(\alpha)}) > 0$ and let*

$$\left(\sum (q_i/p_i)^{1-\alpha} + \sum (p_i/q_i)^\alpha\right)(nVar\,V_n^{(\alpha)})^{-1/2} \to 0 \ \ \textit{for}\ m, n \to \infty. \tag{3.3}$$

Then, under the LD condition (2.5), *as $m, n \to \infty$*

(i) $\mathcal{S}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})/\mathcal{S}_\alpha(\boldsymbol{p}, \boldsymbol{q}) \to 1$ *in probability,*
(ii) $\sqrt{n}(\mathcal{S}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) - \mathcal{S}_\alpha(\boldsymbol{p}, \boldsymbol{q}))/(Var\,V_n^{(\alpha)})^{1/2} \Rightarrow N(0, 1)$,
(iii) $\sqrt{n}\,(\alpha - 1)(\mathcal{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) - \mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}))/\mathcal{CV}(V_n^{(\alpha)}) \Rightarrow N(0, 1)$.

Remark 5 (Plug-in bias) Note that, similarly as in Remark 4, we have $E\mathcal{S}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})/\mathcal{S}_\alpha(\boldsymbol{p}, \boldsymbol{q}) \leq 1$ and, by a similar argument as before, Theorem 2(i) implies

$$E\mathcal{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) - \mathcal{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) \to 0 \quad \text{as} \quad m, n \to \infty.$$

Example 4 (Symmetric divergence for power laws) Consider the symmetric divergence $\mathcal{D}_{1/2}(\boldsymbol{p}, \boldsymbol{q})$ with independent marginals, the case which is often of interest in NGS applications. Note that in this situation, $Var V_n^{(1/2)} = 1/2 - (\sum \sqrt{p_i q_i})^2/2$. Suppose additionally that $p_i = H^{-1}(\beta_1, m)/(i^{\beta_1} l_1(i))$ and $q_i = H^{-1}(\beta_2, m)/(i^{\beta_2} l_2(i))$, $(i = 1, \ldots, m)$ where the notation is as in Example 2 with $0 < \beta_1 \neq \beta_2 < 1$. Then,

$$Var V_n^{(1/2)} \sim \frac{1}{2} - \frac{\sqrt{(1-\beta_1)(1-\beta_2)}}{2 - \beta_1 - \beta_2}$$

and, consequently, (3.3) is seen as equivalent to $m/\sqrt{n} \to 0$ (cf. also Remark 3 above).

With some additional effort, the two CLT results of this section may be extended to degenerate projections. This is discussed in the next section.

3.2 CLTs for degenerate projections

In case of a degenerate projection, the linear term of the power sum Taylor's expansion disappears (cf. formula (B.6) in the appendix) and the condition (3.1) is no longer needed. However, the LD assumption (2.5) has to be slightly strengthened to establish the asymptotic results for the leading (quadratic) term of the appropriate expansion.

3.2.1 Chi-square statistic CLT

The following lemma describing the Chi-square statistic CLT may be of independent interest for models of sparse contingency tables. For a recent discussion of a normal approximation to the Chi-square distribution in such settings, see, e.g., Horgan and Murphy (2013). Here, we apply the Chi-square CLT formulated below to obtain weak limits for the quadratic terms in the entropy and divergence Taylor's expansions leading to Theorems 3 and 4 described in the next subsection. To begin, consider a pair of distributions $(\boldsymbol{p}, \boldsymbol{q})$ and a set of positive weights $\boldsymbol{r} = (r_i)_{i=1}^m$ and define the corresponding Chi-square (χ^2) distance function as

$$\mathcal{X}_{\boldsymbol{r}}^2(\boldsymbol{p}, \boldsymbol{q}) = n \sum \frac{(p_i - q_i)^2}{r_i}.$$

Note that, for instance, the χ^2-distance statistic between the empirical marginals $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})$ is obtained by setting $r_i = p_i + q_i$

$$\mathcal{X}_{\boldsymbol{p}+\boldsymbol{q}}^2(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) = n \sum \frac{(\hat{p}_i - \hat{q}_i)^2}{p_i + q_i}$$

and the Pearson χ^2-statistic is obtained by setting $r_i = p_i$

$$\mathcal{X}_{\boldsymbol{p}}^2(\hat{\boldsymbol{p}}, \boldsymbol{p}) = n \sum \frac{(\hat{p}_i - p_i)^2}{p_i}. \tag{3.4}$$

Below, we denote $\mathcal{X}_{\boldsymbol{u}}^2(\hat{\boldsymbol{u}}, \boldsymbol{u}) =: \mathcal{X}_{\boldsymbol{u}}^2$.

Lemma 2 *Let $(p_{ij})_{i,j=1}^{m}$ be the bivariate distribution of $Z = (X, Y)$ with X and Y having marginals $(p_i)_{i=1}^{m}$ and $(q_i)_{i=1}^{m}$, where $p_i = q_i > 0$. Assume $m \to \infty$ as $n \to \infty$ and*

$$(mn)^{-1} \sum \max \left(p_i^{-1}, p_i^{-2} m^{-1} \right) \to 0. \tag{3.5}$$

Then as $n \to \infty$,

(i) $\dfrac{\mathcal{X}^2_{\boldsymbol{p}}(\hat{\boldsymbol{p}}, \boldsymbol{p}) - m}{\sqrt{2m}} \Rightarrow N(0, 1)$,
and if additionally

$$\sup_n \max_{ij} \frac{p_{ij}}{p_i p_j} = B < \infty \tag{3.6}$$

then also
(ii) $\dfrac{\mathcal{X}^2_{2\boldsymbol{p}}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) - \mu_n}{\sqrt{2}\gamma_n} \Rightarrow N(0, 1)$,
where

$$\begin{aligned} \mu_n &= \sum_i (1 - p_{ii}/p_i) \\ \gamma_n^2 &= \sum_i \frac{(p_i - p_{ii})^2}{p_i^2} + \sum_{1 \le i \ne j \le m} \frac{(p_{ij} + p_{ji})^2}{4 p_i p_j}. \end{aligned} \tag{3.7}$$

Remark 6 Note that for $\mathcal{X}^2_{\boldsymbol{u}}$ the condition (3.5) simplifies to $m/n \to 0$.

Remark 7 Note that under the assumption (3.6), we have $m - 2B \le \gamma_n^2 \le m + B^2$ and therefore $\gamma_n^2 \sim m$. In particular, if $p_{ij} = p_i p_j$ then $\mu_n = \gamma_n^2 = m - 1$.

The proof of the result borrows some ideas from theory of U-statistics (Koroljuk and Borovskich 1994) and may be found in the appendix. Its application is discussed next.

3.2.2 Pure noise and equal marginals CLTs

The first result covers the case of the Renyi entropy when $\boldsymbol{p} := \boldsymbol{u}$. The proof is outlined in the appendix. Recall that for real a and integer k, we define $\binom{a}{k} = a(a-1)\cdots(a-k+1)/k!$

Theorem 3 (Uniform Entropy CLT) *Assume $m \to \infty$ as $n \to \infty$ and $m^2/n = o(n^{-\tau})$ for $\tau > 0$. Then, as $n \to \infty$,*

(i) $\dfrac{n\binom{\alpha}{2}^{-1}[m^{\alpha-1}\mathcal{S}_\alpha(\hat{\boldsymbol{u}})-1]-m}{\sqrt{2m}} \Rightarrow N(0, 1)$,
(ii) $\dfrac{n[\mathcal{H}_\alpha(\hat{\boldsymbol{u}})-\log m-(1-\alpha)^{-1}\log(1+\binom{\alpha}{2}\frac{m}{n})]}{\alpha\sqrt{m/2}} \Rightarrow N(0, 1)$.

Our second CLT result is the following theorem for Renyi divergence when $\boldsymbol{p} = \boldsymbol{q}$. The proof is again deferred to the appendix.

Theorem 4 (Degenerate divergence CLT) *Let $(p_{ij})_{i,j=1}^m$ be the bivariate distribution of $Z = (X, Y)$ with X and Y having marginals $\boldsymbol{p} = (p_i)_{i=1}^m$ and $\boldsymbol{q} = (q_i)_{i=1}^m$, where $p_i = q_i > 0$. Let μ_n and γ_n^2 be given by (3.7). Assume that $m \to \infty$ as $n \to \infty$ and that (3.6) holds, as well as that*

$$\max \left\{ \frac{1}{nm \min p_i^2}, \frac{m}{n \min p_i} \right\} = o(n^{-\tau}). \tag{3.8}$$

Then as $n \to \infty$,

(i) $\frac{n(\alpha(\alpha-1))^{-1}[\mathcal{S}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})-1]-\mu_n}{\sqrt{2}\gamma_n} \Rightarrow N(0, 1)$,

(ii) $\frac{n[\mathcal{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})-(\alpha-1)^{-1}\log(1+\alpha(\alpha-1)\frac{\mu_n}{n})]}{\alpha\sqrt{2}\gamma_n} \Rightarrow N(0, 1)$.

Remark 8 Note that for $\boldsymbol{p} = \boldsymbol{q} = \boldsymbol{u}$, the condition (3.8) reduces to $m^2/n = o(n^{-\tau})$ required in Theorem 3.

3.2.3 Random sample size

When analyzing NGS data, some part of the sequenced reads is frequently removed for technical reasons, for instance, due to poor amplification or reading errors (see next section). In such cases, one effectively deals with a molecular sample of random size. Our CLT results derived earlier may be extended to this case as well, with the help of the following simple result described in Theorem 5 below. Its various versions have been discussed, for instance, in the context of random allocations (see, e.g., Kolchin et al. 1978).

Theorem 5 (Randomized Sample CLT) *Let $(Z_n)_{n=1}^\infty$ be a sequence of bivariate variables supported on an $m_n \times m_n$ integer lattice with distribution $(p_{ij})_{i,j=1}^{m_n}$. Let $(\hat{Z}_n) = (\hat{p}_{ij})_{i,j=1}^{m_n}$ $(n = 1, 2, 3, \ldots,)$ be the sequence of the empirical estimates, each based on an iid sample of (deterministic) size n. Suppose that the statistic $\mathcal{G}_n = \mathcal{G}_n(\hat{p}_{ij})$ satisfies $b_n(\mathcal{G}_n - a_n) \Rightarrow N(0, 1)$ as $n \to \infty$ with some non-random (a_n, b_n). Let $(\nu_n)_{n=1}^\infty$ be a sequence of random variables independent of $(\hat{Z}_n)_{n=1}^\infty$ and following the binomial distributions $bin(n, \tau_n)$ with $0 < \inf_n \tau_n \le \sup_n \tau_n < 1$. Then also,*

$$b_{\nu_n}(\mathcal{G}_{\nu_n} - a_{\nu_n}) \Rightarrow N(0, 1).$$

Proof Denote by $\mathcal{G}_{n_k}$ the random variable $\mathcal{G}_{\nu_k}$ conditional on the event $\nu_k = n_k$ and by Φ the distribution function of the standard normal random variable. By assumption, for any real x we have $P(\mathcal{G}_{n_k} \le x) \to \Phi(x)$, provided that $n_k \to \infty$ as $k \to \infty$. Let $\varepsilon > 0$ be sufficiently small and define $C_\varepsilon(k_0) = \{n_k : k(\tau_k - \varepsilon) \le n_k \le k(\tau_k + \varepsilon), k > k_0\}$. Note that by the weak law of large numbers, $P(\nu_k \in C_\varepsilon(k_0)) \to 1$ as $k_0 \to \infty$.

Therefore,

$$\begin{aligned} P(\mathcal{G}_{\nu_k} \le x, \nu_k \in C_\varepsilon(k_0)) &= \sum_{n_k \in C_\varepsilon(k_0)} P(\mathcal{G}_{n_k} \le x) P(\nu_k = n_k) \\ &= (\Phi(x) + \delta(k_0)) P(\nu_k \in C_\varepsilon(k_0)), \end{aligned}$$

where $\delta(k_0) \to 0$ as $k_0 \to \infty$. Accordingly, as $k_0 \to \infty$ the left-hand side converges to $\lim_k P(\mathcal{G}_{\nu_k} \le x)$ and the right-hand side to $\Phi(x)$ and the result follows.

4 Examples and NGS applications

We start by providing some numerical examples illustrating that, in general, the CLT results discussed above do not hold without assumptions on the relative rate of m and n. Next, we show two examples of applicability of our results to analyzing biodiversity of NGS data. The first one is concerned with comparing the diversity of T-cell receptor populations in transgenic mice, whereas the second one aims at identifying the hepatocellular carcinoma transcription profiles in humans. For the purpose of the T-cell receptors example, we propose a sequential statistical procedure of NGS signal filtering based on our CLT results from the previous sections. We begin by pointing out to some subtleties in the CLT results discussed in Sect. 3.

4.1 Power law and pure noise models

Consider the power law model from Example 2 in Sect. 2.1 with $\beta = 1$ and $l(x) \equiv 1$. Note that in this case $(n \min_i p_i)^{-1} \sim m \log m/n$ as well as $\sum p_i^{\alpha-1}(nVarW_n^{(\alpha)})^{-1/2} \sim O(m(\log^{2\alpha} m/n)^{1/2})$ and therefore the assumptions of Theorem 1 are satisfied as soon as

$$n^{\tau-1} m \to 0 \tag{4.1}$$

for some $\tau > 1/2$. Similarly, the assumption (3.5) of Lemma 2 is satisfied as soon as

$$\log^2 m \frac{m}{n} \to 0. \tag{4.2}$$

In Fig. 1, we illustrate the convergence results of Theorem 1(iii) and Lemma 2(i) for this power law model and $\alpha = 0.5$. The panels of Fig. 1 presents the sample vs standard normal quantile (QQ) plots for the normalized Renyi entropy statistic and the normalized Pearson statistic (3.4) based on $B = 5000$ samples from the power law distribution, each with $m = 1000$ and three different values of $n = m^{1+\varepsilon}$ ($\varepsilon = -0.5, 0.5, 1.5$). As seen from the plots, in absence of (4.1) the CLT result for the Renyi entropy (cf. Theorem 1(iii)) does not hold. Moreover, the middle panel QQ plot indicates that for large m, n satisfying $n = m^{3/2}$ the discrepancy between distribution of the entropy function and its plug-in estimate appears in a form of deterministic shift, indicating the presence of substantial asymptotic bias and hence the lack of

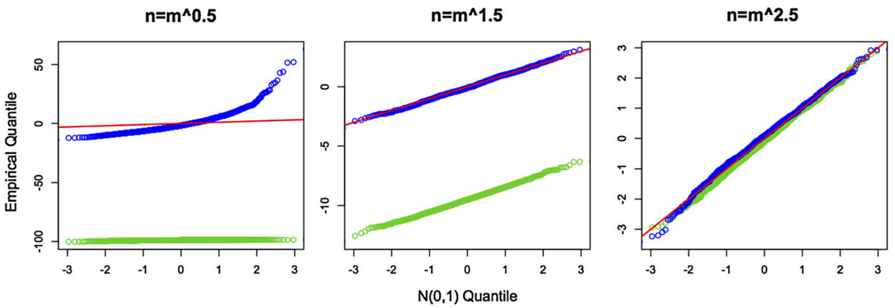

Fig. 1 Projection CLTs. Normal QQ plots for the normalized Renyi entropy (Theorem 1(iii), *lower* (*green*) *curve*) and normalized Pearson χ^2 statistic (Lemma 2(i), *upper*(*blue*) *curve*) for the power law distribution $p_i = 1/i$. The *panels* shows quantile plots with different values of $n = m^{1+\varepsilon}$ ($\varepsilon = -0.5, 0.5, 1.5$) and $m = 1000$. The *solid* (*red*) *line* gives quantiles of the standard normal distribution for reference (Color figure online)

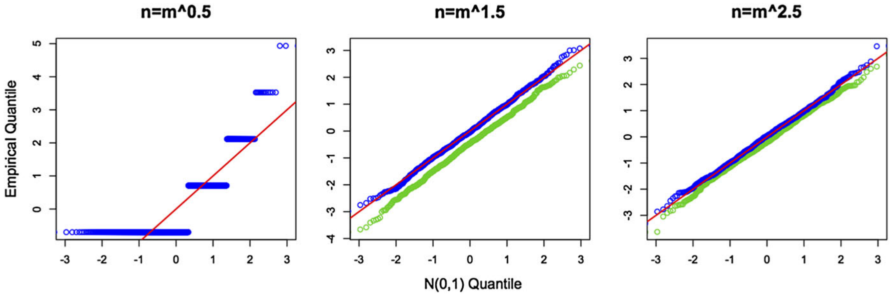

Fig. 2 Degenerate projection CLTs. Normal QQ plots for the normalized uniform Renyi entropy (Theorem 3(ii), represented by the *lower* (*green*) *curve*) and the normalized Pearson χ^2-statistic (Lemma 2(i), represented by the *upper* (*blue*) *curve*) with $p_i = m^{-1}$. The *panels* shows quantile plots with different values of $n = m^{1+\varepsilon}$ ($\varepsilon = -0.5, 0.5, 1.5$) and $m = 1000$. The *solid* (*red*) *line* gives the quantiles of the standard normal distribution for reference. Note that the normalized Renyi entropy is undefined for the first panel (Color figure online)

convergence (3.2). Similarly, when (4.2) is not satisfied, then the Pearson statistic CLT given in Lemma 2(i) fails with the middle panel again indicating that the bias of the estimate does not vanish when m is too large relative to n.

For comparison, we also considered the uniform distribution (pure noise) model $p_i = 1/m$. Note that it may be viewed as a degenerate power law where $\beta = 0$ and $l(x) \equiv 1$. Recall that according to Theorem 3(ii) and Lemma 2(i), the sufficient conditions for the respective CLTs are $m^2/n^{1-\tau} \to 0$ and $m/n \to 0$ (see Remark 6 for the latter one). The necessity of these conditions is illustrated in the panels of Fig. 2, where we again present the (normal) QQ plots for the Renyi ($\alpha = 0.5$) and the Pearson statistics for the same values of B, n and m as in Fig. 1. As seen from these plots, only in the last panel, when $m^2/n \approx 0$, we get good CLT approximation for both statistics. These results appear consistent with our theoretical results from Theorem 3 and Lemma 2.

Although not presented here due to space considerations, similar examples based on the bivariate power laws may be used to illustrate the necessity of the assumptions of type (3.3) and (3.8) in the CLT results for divergence in Theorems 2(iii) and 4(ii).

4.2 Applications to NGS data

Our CLT results described in Sect. 3 were originally motivated by questions rising in NGS data analysis. Below, we describe two examples which adhere to the following basic framework. Denote by $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$ two independent noise distributions each on m support points and assume that a pair $(\boldsymbol{p}, \boldsymbol{q})$ of marginal distributions may be represented as

$$(\boldsymbol{p}, \boldsymbol{q}) = \lambda(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}) + (1-\lambda)(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) \tag{4.3}$$

where $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}})$ is a pair of marginal distributions having no common support points with $(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2)$, and λ is the mixing proportion (or prior probability of signal). We assume that each $\boldsymbol{\varepsilon}$ is a simple finite mixture of K uniform distributions on separate support. Note that the noise-and-signal model from Example 3 in Sect. 2.1 may be viewed as a (univariate) special case of (4.3) with $K = 1$. In the first example below, we took $K = 2$.

Algorithm 1 (*NGS diversity analysis with* $\mathcal{D}_\alpha$ *or* $\mathcal{S}_\alpha$)

(i) *Exponent* (α) *selection*. Use problem-specific criteria (e.g., sample coverage; see Rempala and Seweryn 2013) to identify the appropriate α value. If no prior knowledge exists, the value $\alpha = 1/2$ (the Bhattacharyya distance) may be often used.

(ii) *Noise filtering*. Identify the number of mixture components K and the cutoff count(s) k_m for the support of $\boldsymbol{\varepsilon}_i$ in (4.3) with a sequential (starting from the lowest empirical frequency) procedure based on Lemma 2(i) with $\boldsymbol{p} = \boldsymbol{\varepsilon}_i$ $(i = 1, 2)$. The values of λ is then estimated as the proportion of a sample falling into the m 'noise' categories.

(iii) *Equality testing*. For a predetermined value of α, test the hypothesis $H_0 : \tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$ by comparing the observed value of $\mathcal{D}_\alpha$ (alternatively, $\mathcal{S}_\alpha$) with the asymptotic normal distribution in Theorem 4.

(iv) *Difference quantification*. If H_0 is not rejected, conclude that $\mathcal{D}_\alpha \equiv 0$ ($\mathcal{S}_\alpha \equiv 1$). Otherwise, apply Theorem 2 to obtain confidence bounds for $\mathcal{D}_\alpha$ ($\mathcal{S}_\alpha$).

4.2.1 T-cell receptor populations

In this example, we apply Algorithm 1 to measure the similarity between a pair of T-cell receptor (TCR) populations based on the observed NGS counts of receptor-specific nucleotide sequences. With the current NGS technology, the two main difficulties in comparing TCR populations are to adjust the undersampling bias due to the unobserved rare types and the 'ghost' types created due to the sequencing errors (Wang et al. 2014). The first problem may be often alleviated by applying diversity criteria, like the Renyi entropy and divergence, which allow for the sample-based upweighting

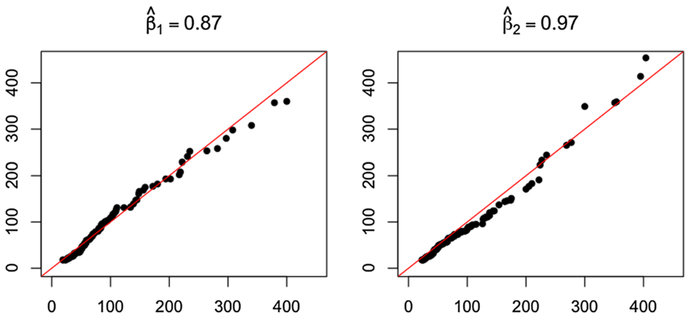

Fig. 3 Power law fit for TCR data. QQ plot of the TCR data against quantiles of a power law distribution with $\beta_1 = 0.87$ ($SE = .05$) and $\beta_2 = 0.97$ ($SE = .05$) values fitted via the least squares method

of rare counts (see Rempala and Seweryn 2013). The second one requires typically additional assumptions, to perform analysis as outlined in Algorithm 1(ii). A recent detailed overview of the TCR diversity analysis methods was presented by Rempala and Seweryn (2013) and earlier on, in a more general context of biodiversity, by Hsieh et al. (2006) and Magurran (2005). For illustration, we analyze two populations derived from the mesenteric lymph nodes (MLN) of a TCR mini-mouse before and after an antibiotic treatment. The details of the experiments and a dataset description are given in Cebula et al. (2013). For the current analysis, it is important to note that, since the experimental groups consisted of different animals, we may consider two experimental groups as independent. The total combined sample size (or sequencing depths) was $n = 72,030$, with initial $m_0 = 6,336$ receptor types. After performing step (ii) of Algorithm 1, $m = 165$ types were identified as "signal" based on the cutoff $k_m = 17$ in both populations. The signal population corresponded to the remaining sample size of 38,896 or about 54 % of the original NGS counts. We used $\mathcal{D}_\alpha$ with $\alpha = 1/2$ as the diversity measure in step (iii)–(iv) of Algorithm 1. Based on Theorem 2, the asymptotic P-value for testing $H_0 : \tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$ was found to be less than 10^{-4} and hence the hypothesis of equal diversity of the two populations was rejected (see Algorithm 1(iii)).

To compare this finding with a more standard parametric analysis, we additionally fitted, with the least squares method, the counts of 165 receptor types in two populations to the power law distributions. Since the respective exponent values for the two fitted populations were found to be different, with $\beta_1 = .87$ (for antibiotic-treated mice) and $\beta_2 = .97$ (for untreated), the parametric analysis confirmed the findings of Algorithm 1. For illustration, the plots of the fitted power law quantiles versus the empirical ones are presented in Fig. 3. Additionally, the diversity of each of the TCR populations in terms of its respective Renyi entropy $\mathcal{H}_{1/2}$ and the Hill number $ENC_{1/2}$ as well as the diversity difference measured by the Renyi divergence $\mathcal{D}_{1/2}$ are listed in Table 1, along with the corresponding asymptotic confidence intervals obtained via Theorems 1 and 2. As seen from the values in Table 1, although the diversity of each of the NGS populations was relatively similar in terms of the two populations count patterns, it differed in terms of the specific TCR types expressed.

Table 1 Results of TCR data analysis

	Antibiotic ($\tilde{\boldsymbol{p}}$)	Control ($\tilde{\boldsymbol{q}}$)
n	39,084	39,084
m	165	165
k_m	17	17
$\hat{\lambda}$	0.46	0.46
$\hat{\beta}$	0.869 (0.05)	0.971 (0.05)
$\mathcal{H}_{1/2}$	4.81 (4.79, 4.82)	4.64 (4.63, 4.67)
$ENC_{1/2}$	122.73 (120.30, 123.97)	103.54 (102.51, 106.70)
$\mathcal{D}_{1/2}$	0.155 (0.147, 0.163)	

The mixture model (4.3) with heavy-tailed power laws fitted to two sets of TCR counts derived from mouse MLN before and after an antibiotic treatment as described in Cebula et al. (2013). The empirical Renyi entropy, the Hill number and the Renyi diversity CIs (in parenthesis) are obtained from the CLT results of Theorems 1 and 2

4.2.2 Gene expression profiling

Beyond Algorithm 1, the results of Sect. 3 may be applied to facilitate various other biodiversity analysis, for instance, in simultaneous comparison of several pairs of molecular samples. We illustrate this with an NGS data example from the recent hepatocellular carcinoma (HCC) study in Chan et al. (2014) which we obtained through the gene expression omnibus (GEO) database. The GEO dataset consists of HCC tumor-infected (T) and healthy liver (N) tissue samples from three individuals denoted below as follows in relation to their original database designations $T1 = HCC448T, T2 = HCC473T, T3 = HCC510T$ and $N1 = HCC448N, N2 = HCC473N, N3 = HCC510N$. For this dataset, one of the questions of research interest was whether the expression profiles of genes associated with regulation of cell proliferation and programmed cell death differ across T and N samples as well as across individuals (cf., e.g., Kong et al. 2013). To address this specific question, in contrast to the previous TCR example, we were thus only interested in a pre-selected subset of the NGS counts. The final values of $m = 1332$ and n between 1.2 and 1.9 million reads[1] were obtained after aligning the pre-selected NGS fragments to the HG19 reference genome with the Tophat2/Bowtie2 software (Kim et al. 2013) and performing the transcript annotation with the Ensembl genome browser (www.ensembl.org). After the final fragments-to-counts conversion, our data analysis was performed in three steps. First, the null hypothesis of the tissue homogeneity $H_0^{all} = \{T_1 = N_1 = T_2 = N_2 = T_3 = N_3\}$ was tested (and rejected) based on the result of Theorem 4 and the corresponding asymptotic p-value obtained from the $\chi^2(3)$ distribution. Next, the hypothesis of the across-individuals homogeneity was tested by evaluating three pairwise null hypothesis $H_0^{ij} = \{\mathcal{D}_{1/2}(T_i, N_i) = \mathcal{D}_{1/2}(T_j, N_j)]\}$, $1 \leq i < j \leq 3$ (each rejected) based on Theorem 4. Finally, having rejected the homogeneity hypothesis, we have used

[1] Based on these values, the empirical versions of the conditions for the relevant theorems in Sect. 3 were considered satisfied.

Table 2 The 95 % confidence intervals for the pairwise symmetric Renyi divergence $\mathcal{D}_{1/2}$ between the tumor and control (healthy) tissues from three individuals based on the profile of expression of pre-selected $m = 1332$ transcripts related to cell proliferation

Hypothesis	Statistic	P-value	$\mathcal{D}_{1/2}$ value (CI)
H_0^{all}	$\sum w_i[\mathcal{D}_{1/2}(i) - \mu_i]^2$	<0.001	NA
$H_0^{1,2}$	$\mathcal{D}_{1/2}(1) - \mathcal{D}_{1/2}(2)$	<0.01	$\mathcal{D}_{1/2}(1)$=0.553 (0.551, 0.555)
$H_0^{2,3}$	$\mathcal{D}_{1/2}(2) - \mathcal{D}_{1/2}(3)$		$\mathcal{D}_{1/2}(2)$=0.292 (0.291, 0.294)
$H_0^{3,1}$	$\mathcal{D}_{1/2}(3) - \mathcal{D}_{1/2}(1)$		$\mathcal{D}_{1/2}(3)$= 0.346 (0.345 0.348)

Here $\mathcal{D}_{1/2}(i)$ denotes $\mathcal{D}_{1/2}(T_i, N_i)$

the result of Theorem 2 to quantify the differences between the three sets of T and N tissue samples. The details of the analysis are presented in Table 2. As seen from the numerical results, it seems that despite the large individual differences between patients, the set of $m = 1332$ genes associated with cell proliferation and death may be used to distinguish between T-type and N-type samples in HCC patients.

5 Summary and Conclusions

We derived two sets of limit theorems for the Renyi entropy and divergence statistics. The first set of results holds for linearalizable statistics (their first-order Taylor approximations exist), whereas the second one holds in the degenerate case (when the first-order approximations vanish) and requires analyzing the quadratic terms in the Taylor expansions. Our Renyi entropy limit theorems complement those obtained elsewhere for the Shannon entropy and divergence.

Based on the CLT results, we have proposed here a new framework for analyzing the diversity of molecular (especially, NGS) data based on the idea of analyzing the frequency/contingency tables where cell counts are highly unbalanced (for instance, as arriving from mixtures of heavy-tailed, power law type and uniform distributions) and the number of cells or, equivalently, the counts distribution support size, m, increases with the sample size n. For analyzing such tables, we suggested using the empirical Renyi entropy and divergence as the statistical measures of, respectively, diversity and pairwise similarity of different molecular sub-populations.

In the two examples of NGS analysis, we have shown how the Renyi entropy methods may be used for filtering out low-frequency noise and for establishing valid confidence bounds in pairwise divergence analysis for pre-selected transcripts. However, it was also seen that to apply our CLT results, the number of transcripts had to be small relative to the sequencing depth. For the special class of heavy-tailed power law distributions, our results in particular indicate that the appropriate entropy CLTs are valid (and thus so is our proposed analysis framework) when, roughly speaking, $m/\sqrt{n} \to 0$ and not otherwise. Such, restriction may be often limiting in very high-diversity NGS data, and other statistics beyond those discussed here and not requiring such condition could be also of interest. We hope to pursue this matter in our future work.

Acknowledgments The authors are grateful to the referees for helpful comments on an earlier version of the manuscript.

References

Agresti A (2002) Categorical Data Analysis, 2nd edn., Wiley series in probability and statisticsWiley, New York

Berkov S, Mutafova B, Christen P (2014) Molecular biodiversity and recent analytical developments: a marriage of convenience. Biotechnol Adv 32(6):1102–1110. doi:10.1016/j.biotechadv.2014.04.005

Campbell AK (2003) Save those molecules: molecular biodiversity and life. J Appl Ecol 40(2):193–203

Cea J (2005) Immunobiology: the immune system in health and disease, 6th edn. Garland Science, New York

Cebula A, Seweryn M, Rempala GA, Pabla SS, McIndoe RA, Denning TL, Bry L, Kraj P, Kisielow P, Ignatowicz L (2013) Thymus-derived regulatory T-cells contribute to tolerance to commensal microbiota. Nature 497(7448):258–262. doi:10.1038/nature12079

Chan THM, Lin CH, Qi L, Fei J, Li Y, Yong KJ, Liu M, Song Y, Chow RKK, Ng VHE, Yuan YF, Tenen DG, Guan XY, Chen L (2014) A disrupted RNA editing balance mediated by adars (adenosine deaminases that act on RNA) in human hepatocellular carcinoma. Gut 63(5):832–843. doi:10.1136/gutjnl-2012-304037

Chao A, Chiu CH, Jost L (2010) Phylogenetic diversity measures based on Hill numbers. Philos Trans R Soc Lond B Biol Sci 365(1558):3599–3609. doi:10.1098/rstb.2010.0272

Chao A, Chiu CH, Hsieh TC (2012) Proposing a resolution to debates on diversity partitioning. Ecology 93(9):2037–2051

Chiu CH, Jost L, Chao A (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecol Monogr 84(1):21–44

de Andrade M, Wang X (2011) Entropy based genetic association tests and gene-gene interaction tests. Stat Appl Genet Mol B. doi:10.2202/1544-6115.1719

Horgan D, Murphy CC (2013) On the convergence of the chi-square and noncentral chi-square distributions to the normal distribution. IEEE Commun 17(12):2233–2237

Hsieh CS, Zheng Y, Liang Y, Fontenot JD, Rudensky AY (2006) An intersection between the self-reactive regulatory and nonregulatory T-cell receptor repertoires. Nat Immunol 7(4):401–410. doi:10.1038/ni1318

Jost L (2007) Partitioning diversity into independent alpha and beta components. Ecology 88(10):2427–2439

Källberg D, Leonenko N, Seleznjev O (2012) Statistical inference for Rényi entropy functionals. In: Conceptual modelling and its theoretical foundations, Springer, New York, pp 36–51

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. doi:10.1186/gb-2013-14-4-r36

Knoblauch A (2008) Closed-form expressions for the moments of the binomial probability distribution. SIAM J Appl Math 69(1):197–204

Kolchin VF, Sevast yanov BA, Chistyakov VP (1978) Random allocations. translated from the Russian. Translation Balakrishnan AV (ed), Scripta series in mathematics. VH Winston & Sons, Washington, DC; distributed by Halsted Press, Wiley , New York-Toronto, Ont-London

Kong D, Chen H, Chen W, Liu S, Wang H, Wu T, Lu H, Kong Q, Huang X, Lu Z (2013) Gene expression profiling analysis of hepatocellular carcinoma. Eur J Med Res 18:44. doi:10.1186/2047-783X-18-44

Koroljuk VS, Borovskich YV (1994) Theory of U-statistics. Mathematics and its applications. Springer, Dordrecht

Krishnamurthy A, Kandasamy K, Poczos B, Wasserman L (2014) Nonparametric estimation of Renyi divergence and friends. In: Proceedings of the 31st international conference on machine learning (ICML 2014), http://research.microsoft.com/apps/pubs/default.aspx?id=256257

Leonenko N, Pronzato L, Savani V et al (2008) A class of Rényi information estimators for multidimensional densities. Ann Stat 36(5):2153–2182 Corrections: Ann. Stat., 2010, 38(6), 3837–3838

Magurran AE (2005) Biological diversity. Curr Biol 15(4):R116–R118. doi:10.1016/j.cub.2005.02.006

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? PLoS Biol 9(8):e1001127. doi:10.1371/journal.pbio.1001127

Nielsen F, Boltz S (2011) The Burbea-Rao and Bhattacharyya centroids. IEEE Trans Inf Theory 57(8):5455–5466

Oh S, Song S, Dasgupta N, Grabowski G (2014) The analytical landscape of static and temporal dynamics in transcriptome data. Front Genet 5:35. doi:10.3389/fgene.2014.00035

O'Rawe JA, Ferson S, Lyon GJ (2015) Accounting for uncertainty in dna sequencing data. Trends Genet. doi:10.1016/j.tig.2014.12.002

Paninski L (2003) Estimation of entropy and mutual information. Neural Comp 15(6):1191–1253

Pardo L (2005) Statistical inference based on divergence measures. CRC Press, Boca Raton

Presley SJ, Scheiner SM, Willig MR (2014) Evaluation of an integrated framework for biodiversity with a new metric for functional dispersion. PLoS One 9(8):e105818. doi:10.1371/journal.pone.0105818

Rempala GA, Seweryn M (2013) Methods for diversity and overlap analysis in t-cell receptor populations. J Math Biol 67(6–7):1339–1368. doi:10.1007/s00285-012-0589-7

Renyi A (1961) On measures of entropy and information. In: 4th Berkeley symposium on mathematical statistics and probability, pp 547–561

Sadee W, Hartmann K, Seweryn M, Pietrzak M, Handelman SK, Rempala GA (2014) Missing heritability of common diseases and treatments outside the protein-coding exome. Hum Genet 133(10):1199–1215. doi:10.1007/s00439-014-1476-7

Shao J (2003) Mathematical Statistics. Springer Texts in Statistics, Springer, New York. http://books.google.com/books?id=cyqTPotl7QcC

Soulier P (2009) Some applications of regular variation in probability and statistics. Escuela Venezolana de Matemáticas. http://evm.ivic.gob.ve/LibroSoulier

Sun W, Hu Y (2013) EQTL mapping using RNA-seq data. Stat Biosci 5(1):198–219. doi:10.1007/s12561-012-9068-3

Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. J Stat Phys 52(1–2):479–487

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. doi:10.1038/nrg2484

Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Łabaj PP, Kreil DP, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian HR, Lancashire LJ, Bessarabova M, Nikolsky Y, Furlanello C, Chierici M, Albanese D, Jurman G, Riccadonna S, Filosi M, Visintainer R, Zhang KK, Li J, Hsieh JH, Svoboda DL, Fuscoe JC, Deng Y, Shi L, Paules RS, Auerbach SS, Tong W (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol 32(9):926–932. doi:10.1038/nbt.3001

Zhang Z, Grabchak M (2014) Nonparametric estimation of Küllback–Leibler divergence. Neural Comput 26(11):2570–2593

Zhang Z, Zhang X (2012) A normal law for the plug-in estimator of entropy. IEEE Trans Inf Theory 58(5):2745–2747