A comparison of computational methods for detecting bursts in neuronal spike trains and their application to human stem cell-derived neuronal networks

Running head: Comparison of burst detectors for spike trains

Ellese Cotterill[1*], Paul Charlesworth[2], Christopher W. Thomas[2], Ole Paulsen[2], Stephen J. Eglen[1]

[1]Cambridge Computational Biology Institute, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK

[2]Department of Physiology, Development and Neuroscience, Physiological Laboratory, Downing Street, Cambridge, CB2 3EG, UK

*Correspondence: Ellese Cotterill, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK. E-mail: ec526@cam.ac.uk.

Conflict of Interest: The authors declare no competing financial interests.

Abbreviations

CMA	Cumulative moving average	MI	MaxInterval
CV	Coefficient of variation	NM	Neural Maintenance
hiPSC	Human induced pluripotent stem cell	PS	Poisson surprise
HSMM	Hidden semi-Markov model	RGC	Retinal ganglion cell
IBI	Interburst interval	RGS	Robust Gaussian surprise
IRT	ISI rank threshold	ROC	Receiver operating characteristic
ISI	Interspike interval	RS	Rank surprise
MCMC	Markov chain Monte Carlo	WAP	Weeks after plating
MEA	Microelectrode array		

ABSTRACT

Accurate identification of bursting activity is an essential element in the characterization of neuronal network activity. Despite this, no one technique for identifying bursts in spike trains has been widely adopted. Instead, many methods have been developed for the analysis of bursting activity, often on an ad hoc basis. Here, we provide an unbiased assessment of the effectiveness of eight of these methods at detecting bursts in a range of spike trains. We suggest a list of features that an ideal burst detection technique should possess, and use synthetic data to assess each method in regards to these properties. We further employ each of the methods to re-analyze microelectrode array (MEA) recordings from mouse retinal ganglion cells, and examine their coherence with bursts detected by a human observer. We show that several common burst detection techniques perform poorly at analyzing spike trains with a variety of properties. We identify four promising burst detection techniques, which are then applied to MEA recordings of networks of human induced pluripotent stem cell (hiPSC)-derived neurons, and used to describe the ontogeny of bursting activity in these networks over several months of development. We conclude that no current method can provide 'perfect' burst detection results across a range of spike trains, however two burst detection techniques, the MaxInterval and logISI methods, outperform compared to others. We provide recommendations for the robust analysis of bursting activity in experimental recordings using current techniques.

NEW & NOTEWORTHY

We provide an unbiased quantitative assessment of eight existing methods for identifying bursts in neuronal spike trains. We reveal limitations in a number of commonly used burst detection techniques and provide recommendations for the best practice for accurate identification of bursts using existing techniques. An analysis of the ontogeny of bursting activity in a novel data set of recordings from hiPSC-derived neuronal networks, using the highest performing burst detectors from our study, is also presented.

Keywords

bursts, spike trains, computational methods, stem cells, development

INTRODUCTION

The tendency of neurons to fire brief periods of spikes in quick succession, or bursts, has been observed extensively both *in vitro* and *in vivo* (Weyand et al., 2001; Pasquale et al., 2010). Bursting is believed to be associated with a variety of physiological processes, such as synapse formation (Maeda et al., 1995) and long-term potentiation (Lisman, 1997). Using recordings of the electrical activity of neurons cultured on microelectrode arrays (MEAs), various aspects of *in vitro* network activity, including bursting, can be readily examined. MEAs have thus been used to study changes in the spontaneous activity patterns exhibited by neuronal networks over development (Wagenaar et al., 2006; Charlesworth et al., 2015). Analysis of bursting activity has also been used as an important tool in applications such as studying the impact of genetic or chemical manipulations on network activity (Eisenman et al., 2015; Charlesworth et al., 2016).

Despite the prevalence of bursting as a feature used to study neuronal network activity, the concept of a burst still lacks a definitive formal definition (Cocatre-Zilgien and Delcomyn, 1992; Gourévitch and Eggermont, 2007) or single widespread technique used for detecting bursts. Instead, a variety of burst detectors exist, many of which have been

developed and verified by researchers on an ad hoc basis using specific data sets and singular experimental conditions.

One common approach to burst detection is to identify periods of bursting using simple thresholds, which impose limits on values such as the minimum firing rate or maximum allowed interspike interval (ISI) in a burst. These thresholds can either be fixed values (Chiappalone et al., 2005; Mukai et al., 2003), or derived from properties of the spike trains, such as the mean ISI (Chen et al., 2009), total spiking rate (Pimashkin et al., 2011) or some form of the distribution of ISIs or discharge density (Cocatre-Zilgien and Delcomyn, 1992; Selinger et al., 2007; Pasquale et al., 2010; Kaneoke and Vitek, 1996; Bakkum et al., 2013; Kapucu et al., 2012). Another type of burst detection techniques are the 'surprise-based' methods, which detect bursts as deviations from an assumed underlying firing rate distribution (Legéndy and Salcman, 1985; Ko et al., 2012; Gourévitch and Eggermont, 2007). There are also a variety of methods based on some variation of these ideas, or which take other approaches entirely (Hennig et al., 2011; Tokdar et al., 2010; Turnbull et al., 2005; Tam, 2002; Xia et al., 2003; Wagenaar et al., 2005; Weihberger et al., 2013).

Most existing studies involving analysis of bursting activity in MEA recordings have been performed on experimental data from rodents (Charlesworth et al., 2015; Mazzoni et al., 2007). In recent years, it has been demonstrated that networks of neurons derived from human stem cells can be grown successfully on MEAs, and exhibit spontaneous electrical activity, including bursting behaviour (Odawara et al., 2016; Heikkilä et al., 2009). These networks often exhibit far more variable bursting activity than more commonly studied rodent neuronal networks (Kapucu et al., 2012). There is thus a demand for the development of robust standardized analysis methods for identifying bursts in such networks.

Here, we have reviewed eight existing burst detectors, selected to encompass a range of contemporary burst analysis methods, and evaluated their effectiveness at detecting bursts, particularly in spike trains with properties resembling those of human stem cell-derived neuronal networks. Each burst detector was used to analyze bursts in synthetic

spike trains and *in vitro* MEA recordings from mouse retinal ganglion cells (RGCs). This allowed for a quantitative assessment of the performance of each method in a variety of contexts. Based on these results, we offer suggestions to researchers regarding the best approaches for comprehensive burst analysis. The highest performing methods in our study were also used to describe the ontogeny of bursting activity in networks of human induced pluripotent stem cell (hiPSC)-derived neurons over several months of development.

MATERIALS AND METHODS

Burst analysis methods

Eight burst detectors that we believed to be representative of the major approaches to burst detection and that have sufficient general applicability to allow for their use on a variety of spike trains were chosen for analysis. Other methods were excluded from the analysis for a variety of reasons, including that they do not explicitly identify the location of bursts in a spike train (van Elburg and van Ooyen, 2004), or we believe that they have been superseded by more refined methods (Selinger et al., 2007; Chiappalone et al., 2005).

A brief description of each of the eight burst detectors applied to a single spike train is given below, and we refer the reader to the original sources for detailed descriptions. Where possible, we reused existing code from the original authors to implement each method. All analyses presented here were performed using R statistical software (R Core Team, 2015), and the code used to implement each burst detector is publicly available at `https://github.com/ellesec/burstanalysis`.

In the implementation of each method, the minimum number of spikes in a burst was set to three and other parameters were left set to the standard parameters suggested by the authors (Table 1). The exception to this was the three surprise-based burst detectors, for which the minimum surprise value was set to $-\log(0.01)$ for all three methods for

consistency.

LogISI method (Pasquale et al., 2010)

Bursts are detected using the histogram of the log adjusted interspike intervals (ISIs) on a spike train. The peaks of this histogram are found using a tailor-made peak finding algorithm outlined in Pasquale et al. (2010), and the largest peak corresponding to an ISI of 100 ms or less is set as the intra-burst peak. In the absence of such a peak, no bursts are found. The minimum values between the intra-burst peak and all subsequent peaks are found, and a void parameter, which represents how well the peaks are separated, is calculated for each minima. The ISI value corresponding to the first minimum at which the void parameter exceeds a threshold value of 0.7 is set as the cutoff value for burst detection, $maxISI$. Bursts are then detected as any series of three or more spikes separated by ISIs smaller than $maxISI$. If no cutoff is found, or if $maxISI > 100$ ms, bursts are found using a 100 ms cutoff, and then extended to include any spikes within $maxISI$ of the edges of each burst.

Cumulative Moving Average (CMA) method (Kapucu et al., 2012)

The cumulative moving average (CMA) of the histogram of ISIs is calculated. The skewness of this CMA distribution is used to determine the values of two parameters, α_1 and α_2, according to the scale given in Kapucu et al. (2012) and shown in Table 1. The ISI value of the histogram bin at which the CMA is closest in value to $\alpha_1 \cdot CMA_{MAX}$ is set as $maxISI$, where CMA_{MAX} is the peak of the CMA distribution. Again, bursts are defined as sequences of more than two spikes separated by ISIs less than $maxISI$.

Kapucu et al. (2012) also suggest expanding these bursts to include burst related spikes, which are found using a cutoff set at the histogram bin at which the CMA is closest to $\alpha_2 \cdot CMA_{MAX}$. Any spikes within this cutoff distance from the beginning or end of the original bursts are classified as burst related spikes. However, for our purposes, we only examined the original burst cores detected from this method, and omitted any burst related spikes.

ISI Rank Threshold (IRT) method (Hennig et al., 2011)

The rank, $R(t)$, of each ISI relative to the largest ISI on the spike train is calculated. A rank cutoff, θ_R, is chosen, and a spike count cutoff, θ_C, is calculated from the distribution of spike counts over one second intervals on the spike train. A burst begins at a time t if the spike count over the following second exceeds θ_C, and its subsequent ISI satisfies $R(t) < \theta_R$. The burst continues until the spike count over the following one second interval falls below $\frac{\theta_C}{2}$.

Poisson Surprise (PS) method (Legéndy and Salcman, 1985)

The baseline firing rate on a spike train is assumed to follow a Poisson process with rate λ equal to the mean firing rate over the entire train. The Poisson surprise statistic for an interval of length T containing N spikes is defined as

$$S = -\log(p)$$

where p is the probability of N or more spikes randomly occurring in an interval of length T in the underlying Poisson process. Bursts are chosen so as to maximize the Poisson surprise statistic over the entire spike train using a surprise maximization algorithm outlined in Legéndy and Salcman (1985), and any bursts with a Poisson surprise value below a threshold significance level are discarded.

Rank Surprise (RS) method (Gourévitch and Eggermont, 2007)

The ISIs on the spike train are ranked, with the smallest ISI given a rank of 1. For each possible bursting interval, the rank of all ISIs on the interval are summed, and the probability, p, of a value of equal or lesser value being drawn randomly from a discrete uniform sum distribution is calculated. Bursts are chosen so as to maximize the rank surprise statistic, defined as $RS = -\log(p)$, across the entire spike train, and any bursts with a rank surprise statistic below a pre-defined significance threshold are discarded.

Robust Gaussian Surprise (RGS) method (Ko et al., 2012)

Bursts are regarded as outliers from a "central distribution" of ISIs, which is estimated from the distribution of normalized $\log(ISIs)$ on a spike train. ISIs are considered to be potentially within bursts if they lie below -2.58 times the median absolute deviation of this distribution. For each potential burst, the Robust Gaussian Burst Surprise value, $GS_B = -\log(p)$, is calculated, where p is the probability that the sum of the normalized $\log(ISIs)$ in the interval is less than or equal to the sum of an equivalent number of i.i.d. Gaussian random variables with mean and standard deviation equal to those of the central distribution. These initial bursts are then extended to include surrounding spikes until the maximal surprise value is found. Finally, any bursts with a surprise value below a pre-defined significance threshold are discarded.

Hidden Semi-Markov Model (HSMM) method (Tokdar et al., 2010)

Neurons are assumed to stochastically alternate between a "non-bursting" and "bursting" state, labelled states 0 and 1 respectively. Spiking activity is modeled using a hidden semi-Markov model, with transition times between the two states modeled as two Gamma distributions, f_0^{ITI} and f_1^{ITI}. Within each of the two states, the distribution of ISIs are modeled using two additional Gamma distributions, f_0^{ISI} and f_1^{ISI}. The parameters of these four distributions are learned from the data. Under these assumptions, the posterior probability that the neuron is in a bursting state at any given time is calculated using a Markov chain Monte Carlo (MCMC) method, and any periods in which this probability exceeds a threshold value are classified as bursts. An R package to implement this method is available at `https://stat.duke.edu/~st118/Software/`.

MaxInterval (MI) method (Nex Technologies, 2014)

Any series of consecutive spikes fulfilling five threshold parameters, the values of which are chosen by the user, are classified as bursts. For our purposes, the values of the parameters were those specified in the NeuroExplorer Manual (Nex Technologies, 2014), see Table 1.

When applied to data sets consisting of multiple spike trains, for example multiple chan-

nels from a single MEA recording, most burst detectors analyze each spike train individually, calculating any associated parameter values, e.g.$maxISI$, separately for each spike train. The exceptions to this are the MI method, which uses the same fixed parameters to detect bursts on all electrodes, and the RGS method, which combines the ISIs from all channels and uses this pooled data set to determine the characteristics of the "central distribution" and find the initial bursting periods on each electrode.

Analysis of synthetic data

The performance of each method was evaluated against a list of properties that we deemed desirable in a burst detector, shown in Table 2. For properties D1–D3, performance was based on the details of the method's implementation, while for the remaining properties (D4–D11), testing on simulated data was performed. Simulated data was used for this purpose because it allowed us to generate spike trains with specific properties of interest. By explicitly generating periods of bursting activity in these spike trains, we were also able to compare the results of each burst detector with the 'ground truth' bursting behaviour. Simulated spike trains were produced using the models outlined below, with the parameter values specified in Table 3.

Poisson and Gamma distributions

Two types of non-bursting spike trains were simulated, one with Poisson distributed ISIs, and the other with Gamma distributed ISIs. The smallest 10th percentile of ISIs were removed from each spike train by omitting the corresponding spikes, to eliminate any burst-like behaviour arising randomly in the simulated data.

Inhomogeneous Poisson distribution

Spike trains with non-stationary firing rates and no bursts were simulated using a Poisson process with non-homogeneous intensity, $\lambda(t)$. To eliminate any possible bursting behaviour, spikes corresponding to the smallest 10th percentile of ISIs were removed from each spike train.

Poisson Bursting

Bursting spike trains were simulated using the Poisson bursting model. The location of the center of each burst on a spike train was modeled using a Poisson process with a fixed rate, λ. The number of spikes in each of the bursts was drawn from a Poisson distribution with mean n. The position of the spikes in each burst relative to the burst center were drawn from a uniform distribution with range r and mean 0. Where two bursts overlapped, only the first was kept.

To simulate spike trains with non-stationarity in their bursting properties, the values of n and r were drawn randomly from a uniform distribution for each burst, rather than being held as fixed values. Only the resulting bursts with within-burst firing rate above 5 Hz were retained.

To simulate noise in bursting spike trains, noise spikes were modeled with Gamma distributed ISIs with the smallest 10th percentile of ISIs removed. These noise spikes were added to the Poisson bursting spike train and any noise spikes within 0.5 s of the limits of each burst were removed, to prevent any overlap between burst and noise spikes.

For each desirable property, one hundred spike trains of duration 300 s were simulated and analyzed using each of the burst detectors detailed above. Examples of the simulated spike trains used for evaluating each desirable property are shown in Figure 1. A comparison of the 'ground truth' bursting activity and the results from each burst detector was then performed. For spike trains containing both bursts and noise spikes, this involved examining the fraction of true positive spikes, defined as the proportion of within-burst spikes correctly identified as being in bursts, and the fraction of false positive spikes, defined as the fraction of all noise spikes erroneously identified as being within bursts, found by each burst detector.

Analysis of mouse RGC data

MEA recordings of mouse retinal ganglion cells (RGCs) from Demas et al. (2003) were re-analyzed using the burst detectors in our study. These MEA recordings are available from Eglen et al. (2014). Four hour-long recordings of control mice retina at postnatal days 9, 11, 13 and 15 were chosen for re-analysis. For spike trains from five randomly chosen electrodes from each recording, bursts were annotated by visual inspection by one of the authors (EC). Figure 2 shows examples of annotated bursts for spike trains at each age. As MEA recordings do not provide the 'ground truth' location of bursts, these visually identified bursts were taken as a proxy for 'ground truth' bursts and used to compare the results from each burst detector against. Comparison to visually identified bursts has been used previously to assess burst detection techniques (Chen et al., 2009; Gourévitch and Eggermont, 2007; Pasquale et al., 2010). For each burst detector in our study, bursts were detected on the annotated spike trains using a variety of input parameters, and the sensitivity and specificity of each method examined.

To assess their robustness, we chose a key parameter to vary for each burst detector. The parameter that was varied to examine the sensitivity and specificity of the HSMM and surprise-based methods was the probability cutoff, while for the IRT method, the spike count cutoff, θ_C, was altered. For the logISI method, the limit on the maximum allowed ISI cutoff value was varied from its initial value of 100 ms. For the MI method, most parameter values shown in Table 1 were maintained, excluding the maximum beginning and end ISIs, which were varied so that the maximum end ISI was always 0.130 s greater than the maximum beginning ISI. Finally, for the CMA method, for which there were no obvious parameters to vary, only a single value for sensitivity and specificity was found.

Receiver operating characteristic (ROC) curves were produced that plotted 1-specificity versus sensitivity for various parameter values. Sensitivity was defined as the number of spikes correctly detected as being within bursts, as a fraction of the total number of spikes in the visually annotated bursts. The value of 1-specificity, or the false positive rate, was the number of spikes that were falsely detected as being within bursts, as a fraction of the

total number of spikes that were not a part of the 'ground truth' bursts.

Experimental details for hiPSC-derived neural network recordings

Neuronal networks were grown from late stage neuronal precursors differentiated from human induced pluripotent stem cells (hiPSCs) (Axol Bioscience, Moneta Building, Babraham Research Campus, Cambridge). hiPSCs were generated by reprogramming of embryonic cord blood cells and then differentiated to the neural lineage using protocols based on those in Shi et al. (2012). All of the recordings (447 recordings from 73 MEA platings on 11 plating dates, 4 thawed vials) were obtained using a single line and neural induction (AX0015, `http://www.axolbio.com/page/neural-stem-cells-cerebral-cortex`).

Late stage neuronal precursors (1×10^6) were thawed and expanded by growing on 6 well tissue culture plates coated with polyornithine and laminin (2×10^5 cells/well) in 'Neural Maintenance' (NM) medium, supplemented with 10 μM Y-27632 (rho-associated protein kinase inhibitor) for the first 24 hours. After 4–5 days cells were dissociated with Accutase, centrifuged, resuspended in NM and either plated to MEAs ($2 \times 10^4 - 1 \times 10^5$), or expanded further on six-well plates as above. MEAs (60MEA200/30-Ti, Multi Channel Systems, Reutlingen, Germany) were coated with polylysine followed by laminin as described previously (Charlesworth et al., 2016).

hiPSN–MEA cultures were maintained in NM medium under zero evaporation lids (Potter and DeMarse, 2001) housed in tissue culture incubators maintained humidified at 37°C and 5% CO_2 / 10% O_2 / 85% N_2. Media was completely exchanged after 24 hours to remove Y-27632. Thereafter, MEA cultures were fed by exchanging 40–50% medium with fresh NM three times per week. NM media composition was a 1:1 mixture of N2 and B27 supplemented media. N2 media: DMEM/F12 + N2 supplement and 5 μg ml^{-1} insulin, 1mM l-glutamine, 100 μM nonessential amino acids. B27 media: Neurobasal + B27 supplement and 0.5 mM l-glutamine.

Recordings (300 s) of spontaneous extracellular neuronal activity in hiPSN-MEA cultures

were made weekly using an MEA system supplied by Multi Channel Systems (MEA 1060INV, with 60MEA200/30-Ti arrays; titanium nitride electrodes, 30 μm diameter, 200 μm spacing, internal reference electrode). The signal was sampled at 25 kHz and stored using a 64-channel data acquisition board (MC Card; Multi Channel Systems) and the acquisition software MCRack (Multi Channel Systems). Action potentials were detected by crossing of a threshold set to a level of 6 standard deviations from the baseline noise level. Record samples (1 ms pre- and 2 ms post-crossing of threshold) confirmed the characteristic action potential waveform. Action potential timestamps were extracted to text file using batch scripts written for NeuroExplorer (Nex Technologies, Littleton, MA). Recordings made at dates above sixteen weeks after plating were excluded from the analysis due to the small number of data points, resulting in 424 recordings being analyzed. All experiments using human stem cells were vetted and approved by the Steering Committee for the UK Stem Cell Bank and for the Import of Stem Cell Lines in 2012. All procedures were compliant with the UK Code of Practice for the Use of Human Stem Cell Lines.

RESULTS

Desirable properties for a burst detector

To evaluate the performance of each burst detector, the methods were assessed against eleven desirable properties, listed in Table 2. The optimal burst detector would ideally possess all of these desirable properties. For binary properties, D1–D4, each method was judged to either possess the property or not, while for properties D5–D11, the performance of each method was ranked against the other methods, based on the median and variance of its performance at analyzing one hundred synthetic spike trains.

The first desirable property of a burst detector was that it was deterministic (D1), as this ensures reproducibility and removes the need to find a 'consensus' set of bursts across repeated trials. The only non-deterministic burst detector was the HSMM method, due to its use of MCMC methods. The bursts detected by this method varied considerably between

trials. For example, when used repeatedly to analyse one simulated 300 s Poisson bursting spike train with burst frequency of 0.2 Hz, the HSMM method identified 51 ± 9.75 bursts (mean $\pm$ s.d.) over one hundred trials.

Another desirable property for the burst detectors was that they did not assume that ISIs follow a specific statistical distribution (D2). There is no consensus on which type of statistical distribution best represents underlying spike train activity, and any assumption that this activity follows a fixed statistical distribution restricts the applicability of a method to a narrow range of spike trains. Most methods do not assume a fixed statistical distribution for the underlying spike train, excluding the PS, RGS and HSMM methods, which assume that ISIs can be modeled using a Poisson process, Gaussian distribution and Gamma distributions respectively. However, the PS and RGS methods detect bursts as periods of deviation from these underlying firing rate distributions. These methods thus remain somewhat robust when the distribution assumptions are not met, as 'surprising' sequences of spikes as measured by one distribution will generally also correspond to high surprise values from other distributions commonly used to model spike trains (Legéndy and Salcman, 1985).

A common issue that arises when applying burst detection techniques to large sets of spike trains that have high variability in their statistical properties, such as those from MEA recordings of human neuronal networks, is how to accurately choose the parameters for burst detection. This is further confounded when burst detectors are used to analyze MEA recordings spanning a large range of developmental ages, or differing experimental conditions. Thus, ideally, a burst detector should have few parameters (D3), to minimize the impact of how parameter values affect the resultant detected bursts. Most methods in our study only required one or two parameters. The MI method, however, required five parameters to implement burst detection. The HSMM method also required a large number (N=23) of parameters, however, many of these are initial values that are later optimized by the algorithm, and can be left set to the values suggested by the original authors with little impact on the effectiveness of the method.

With the increasing prevalence of high density MEAs that contain up to several thousand electrodes (Maccione et al., 2014), as well as the use of multi-well MEAs in applications such as high-throughput neurotoxicology screening (Valdivia et al., 2014; Nicolas et al., 2014) and drug safety testing (Gilchrist et al., 2015), the computational complexity of each method must also be considered. To assess computational time (D4), each method was used to analyze one hundred simulated spike trains of five minutes duration with average firing rate of 1 Hz. Most methods required on average only a fraction of a second to analyze each spike train using a standard personal computer. The exception to this was the HSMM method, which had an average computational time more than 20 times greater than any other method.

A common feature seen in MEA recordings of human neuronal networks is many electrodes that record sparse or no bursting behaviour. An ideal burst detector would find no or very little bursting activity in these spike trains. Most burst detectors performed reasonably well at detecting a low amount of bursting activity in spike trains simulated to exhibit an absence of bursting behaviour (D5). The major exception to this was the HSMM method, which had a tendency to significantly overestimate bursting behaviour in these spike trains (Figure 3A).

When a non-stationary firing rate was incorporated into non-bursting spike trains (D6), the number of erroneous bursts detected by most methods increased (Figure 3B). The PS and CMA methods, in particular, showed a significant increase in the amount of bursting activity detected in non-stationary spike trains, compared to those with a static mean firing rate. These methods tend to detect periods of 'unusual' activity as bursts, and thus showed a tendency to detect bursts in the regions of relatively high firing rate in these spike trains.

An ideal burst detector should also detect bursts accurately in spike trains that contain only bursting activity, especially those in which the bursts are regular and well separated

(D7). Most methods possessed this property, and could identify over 90% of the spikes within bursts in simulated spike trains containing regular bursting behaviour (Figure 3C). The exceptions to this were the RS, IRT and RGS methods, which consistently detected less than half of the bursts in these synthetic spike trains. This result is unsurprising, since these three methods use thresholds that impose a limit on the maximum proportion of ISIs in a spike train that can be classified as being within bursts.

We also analyzed the performance of each burst detector on simulated spike trains with less standard bursting behaviour. This included spike trains containing non-stationary bursting activity (D8), in the form of bursts with variable lengths and durations. The logISI, HSMM, PS and MI methods correctly identified most spikes in bursts in these spike trains (Figure 3D). The fraction of bursting spikes detected by the CMA method varied considerably across the one hundred simulated spike trains, and it usually correctly identified a significantly lower proportion of within-burst spikes in these spike trains, compared to those containing regular bursting activity. The RS, IRT and RGS methods continued to detect only a small proportion of the bursting activity.

We also examined the performance of each burst detector on spike trains containing bursts with long durations and relatively low within-burst firing rates (D9). For these spike trains, only the PS and HSMM methods gave reasonably accurate results for both the fraction of spikes in bursts and the number of bursts in the spike trains (Figure 4A,B). The MI and CMA methods both correctly allocated a large proportion of the spikes as being within bursts, but tended to separate the long bursts into shorter, more frequent bursts, while the remaining methods greatly underestimated the prevalence of bursting activity in the simulated data (Figure 4A,B).

Another type of non-standard bursting activity seen in human network recordings is the presence of short, poorly separated bursts occurring at a high frequency. When used to analyze spike trains with very frequent bursting behaviour (D10), the MI, logISI and HSMM methods could correctly identify the majority of spikes as being within bursts, but had

a tendency to combine the short bursts into a smaller number of bursts with longer durations (Figure 4C,D). The CMA method most accurately detected the large number of bursts in these high frequency spike trains, but tended to underestimate the proportion of spikes in bursts. The RS, IRT and RGS methods were only able to identify a low fraction of bursting spikes and low number of bursts in these spike trains (Figure 4C,D).

Finally, an ideal burst detector should correctly differentiate between bursting and non-bursting periods in spike trains in which some spiking activity occurs outside of bursts (D11). By comparing each method's output to the ground truth bursting behaviour in simulated spike trains containing both bursts and noise, we examined the fraction of correctly identified within-burst spikes, as well as the fraction of noise spikes erroneously detected as being within bursts. The MI, CMA, logISI and HSMM methods displayed reasonably high true positive rates for identifying bursting spikes, however, of these, the logISI and HSMM methods tended to classify a higher proportion of noise spikes as being within bursts (Figure 4E,F). The RS, IRT and RGS methods exhibited very low false positive rates, but this came at the expense of quite low true positive rates, giving them low overall recall. The performance of the PS method was between these two extremes, with both lower true positive and false positive rates than the highest performing burst detectors.

Tables 4 and 5 summarize the performance of each burst detector across all desirable properties. The ranking in Table 5 was based on the median and range of the boxplots in Figures 3 and 4, with methods with similar results given equal rankings. Three methods clearly underperformed based on this ranking, namely the RS, IRT and RGS methods. To further assess the performance of the burst detectors, they were each used to analyze bursting activity in experimental recordings from mouse RGCs.

Preliminary analysis of mouse RGC data

MEA recordings of mouse RGCs from Demas et al. (2003), a study which examined the developmental changes in spontaneous retinal activity in normal and dark-reared mice,

were re-analyzed using our burst detectors. The sensitivity and specificity of each method at a range of parameter values was calculated, and averaged across the five annotated spike trains from each recording, to produce the ROC curves in Figure 5. The ROC curves for P11 and P13 are ommitted, as they resemble the results at P9. Because of innate restrictions on how bursts are defined by each method, for example that bursts must contain a minimum of three spikes, many burst detectors did not allocate either no spikes or 100% of spikes as being within bursts for any choice of parameter values, and thus the ROC curves do not span the entire range of sensitivity and specificity values. The methods were thus assessed by their minimum distance from the point of perfect classification at (0,1), rather than the area under the ROC curve.

The MI method exhibited strong performance across all ages, and reached very high levels of sensitivity and specificity for a specific choice of parameter values at each age. The logISI, PS and CMA methods also had promising performance across most ages, however at P15 the PS and CMA methods exhibited higher false positive rates (Figure 5B). The results from the RGS method did not vary significantly as its parameter value was changed, and it was unable to reach high levels of sensitivity for any choice of parameter values. The sensitivity and specificity of the RS and IRT methods, on the other hand, spanned a range of values, however these methods did not reach the levels of sensitivity of the other methods. The HSMM method reached high sensitivity levels, but only at the cost of low specificity, and generally performed poorly across all ages.

Evaluation of the methods

Based on the assessment of the burst detectors against the desirable properties, the RS, IRT and RGS methods underperformed compared to the other methods. This was reinforced through their performance when compared to visually annotated bursts in mouse RGC recordings, where they did not reach the levels of sensitivity of the other methods in our study. These three methods were thus eliminated from further analysis. The HSMM method had average performance on simulated data (D5–D11), however its complex implementation meant that it was the lowest performing method on properties D1–D4. The

high false positive rate of the HSMM method across all ages when analyzing experimental data cemented our decision to exclude this method from further consideration.

Further analysis of mouse RGC data

The remaining four methods (PS, MI, CMA and logISI) were used to analyze the complete set of spike trains from each of the four control mouse RGC recordings from Demas et al. (2003). In this case, the parameters used for the analysis were based on those that resulted in the best performance in the ROC curves, as measured by the distance of the curve from the point of perfect classification in the top-left corner. In the original report, rather than explicitly identifying the location of bursts, Demas et al. (2003) used the autocorrelation of each spike train to determine the average burst duration at each age. By explicitly identifying bursts using our four burst detectors, we were able to provide a more detailed description of the bursting activity and compare this with the authors' original results.

The four burst detectors were generally in agreement about the proportion of spikes in bursts across all ages, and showed a decrease in the fraction of spikes in bursts with increasing developmental age (Figure 6A). This concurs with the analysis of the original authors, who found that only very few spikes occurred outside of bursts at early ages, while by P15, many cells were active outside of bursts (Demas et al., 2003).

In terms of burst duration, all four methods showed very similar results over P9 to P13, which resemble the values found by the autocorrelogram method. However, at P15 there was a significant deviation between the results of the PS method and the other burst detectors. The MI, CMA and logISI methods followed the trend of decreasing burst duration with age, as in the original analysis, while the PS method detected a significant increase in burst duration at P15. This can be attributed to the fact that at P15 many spike trains exhibited regular 'bursting episodes', periods of high activity generally spanning 10–20 s that consisted of a series of shorter bursts. The PS method had a tendency to classify these bursting episodes as one long burst, while the other methods generally broke up these periods into several shorter bursts, as shown in the example spike train in Figure 6D.

Analysis of hiPSC-derived neuronal network recordings

To further assess these methods, 424 MEA recordings from 73 hiPSC-derived neuronal networks recorded at regular intervals from two to sixteen weeks after plating (WAP) were analyzed using the four best burst detectors. The MI, PS and logISI parameters used for this analysis were chosen by inspection as those that most accurately detected bursts on five randomly chosen spike trains with mean firing rate close to 1 Hz, which is the minimum firing rate at which regular bursting activity tends to arise. For the CMA method, the scale for α_1 values from Kapucu et al. (2012) was used, and the authors' suggestions for post hoc screening were employed, with any spike trains that were found to have average burst duration greater than 5 seconds or an average burst length above 50 spikes per burst, declared as non-bursting. The resultant parameters used to implement each method are shown in Table 6.

Although there were some differences in the absolute level of bursting activity detected by the different analysis methods, the results from most methods suggested a general trend of 'ramping up' of bursting, in terms of fraction of spikes in bursts, with increasing developmental age (Figure 7B). In general, however, the prevalence of bursting activity in these human network recordings tended to be significantly lower than that commonly seen in recordings of rat and mouse hippocampal or cortical networks (Charlesworth et al., 2015; Chiappalone et al., 2005). The results also suggest that the prevalence of bursting activity in these networks may decrease with age after reaching a peak around 14 WAP (Figure 7B). This would be consistent with previous studies using calcium imaging of human pluripotent stem cell-derived neuronal networks, which found that bursting activity decreases at later stages of development, when it is replaced by more complex firing patterns (Kirwan et al., 2015). However, additional recordings at later time points would be required to confirm this trend in our data.

Unlike some previous studies of rodent neuronal networks (Charlesworth et al., 2015; Demas et al., 2003), there was no obvious relationship between burst duration and culture age in our hiPSC-derived network recordings, with bursts remaining short over the entire

developmental period. Similarly, the degree of regularity of the bursting activity, captured by the coefficient of variation of interburst intervals (CV of IBI), did not appear to change significantly with increasing developmental age (Figure 7A, C).

To quantify the differences between the bursts found by each burst detector in these recordings, we converted each spike train to a time series by dividing the 300 s recording period into 50 ms bins. A binary vector was then found for each burst detector, which took a value of one if the spike train was found to be in a bursting state during that time bin, or zero otherwise. The Hamming distance, which represents the number of points at which two binary strings differ, was calculated between each pair of methods for every spike train on which bursting activity was detected by all four methods, and normalized to represent the fraction of time bins in which the results from each pair of burst detectors differed.

Figure 7 shows that the median Hamming distance between most of the methods was below 5% at most WAPs. At WAP 12, however, there was a peak in Hamming distances, in particular those measuring the difference between the bursts detected by the MI method and other burst detectors. At 12 WAP, the recordings on average exhibited a higher mean firing rate and lower variability of ISIs compared to recordings at other time points, with many electrodes recording tonic spiking or bursting activity at a high frequency (e.g. Figure 8 E,F). As the MI method detects bursts based on the absolute length of ISIs, this method had a tendency to find a large proportion of bursting activity in these high frequency spike trains, while the other methods, which detect bursts as periods of high firing rate relative to the background activity, generally detected a much lower proportion of bursting activity in these spike trains. Altering the MI method parameters at this WAP, to reduce the maximum allowed beginning and end ISIs in a burst, could bring its results more in line with those of the other burst detectors.

Visual inspection of spike trains at other WAPs was also performed to gather insight on the differences between the bursts detected by each method at these ages (Figure 8). In

several examples, the CMA method failed to detect numerous periods that visual inspection and the other burst detectors generally classified as bursting (Figure 8B, C). This may account for the lower proportion of spikes in bursts found by this method, compared to the other burst detectors (Figure 7B). The logISI method also detected low proportions of spikes in bursts across many WAPs (Figure 7B). This may explain the generally low Hamming distances between the CMA and logISI methods (Figure 7D). Additionally, the PS method tended to combine bursts that other methods detected as separate bursts, and extend bursts to incorporate additional spikes that visual inspection would suggest should not be included in bursts, accounting for the longer burst durations found by this method (Figure 7A). Although no method agreed perfectly with how we would assign bursts in these recordings of hiPSC-derived neuronal networks, when a large subset of spike trains were visually examined, out of the four methods examined here, on average the MI method corresponded most closely to how we would annotate bursts visually.

DISCUSSION

Despite the important role of accurate burst detection in analyzing neuronal network activity in a variety of contexts, a consistently widely used method for burst analysis is yet to be adopted. By examining the performance of eight burst detectors at analyzing both synthetic and experimental data, we found that a number of existing methods perform poorly at identifying bursts in spike trains with a variety of properties. We identified four burst detectors that outperformed compared to other existing methods, and used these to analyze bursting activity in recordings of hiPSC-derived neuronal networks over several months of development.

We have shown that a number of burst detectors that were developed based on recordings from single experimental conditions do not necessarily generalize to use on other types of spike trains. For example, the RGS method, which was originally developed to analyze dopaminergic neurons, could not detect the majority of bursts in simulated spike trains, and also performed poorly at analyzing experimental recordings from mouse RGCs, even

when its probabilistic threshold parameter was varied over a large range. Other studies have also found issues using the RGS method to analyze changes in bursting behaviour under different drug effects (Eisenman et al., 2015).

The IRT method also performed poorly at detecting bursts in a range of different spike trains. Unlike the other methods included in our study, this method was not published in a methods paper, but rather was a heuristic method designed for the analysis of a specific data set which was not spike sorted (Hennig et al., 2011), so its lack of adaptability is not surprising.

We have also shown that the complexity of a burst detector does not necessarily correlate with its effectiveness. The most complex method in our study, the HSMM method, often performed only equally well or worse than simpler methods, particularly in non-bursting conditions. Furthermore, the high computational time and non-deterministic nature of this method severely limits its ability to be scaled up for use in high-throughput analysis of MEAs, which is becoming increasing prevalent in applications such as large-scale neurotoxicity testing (Nicolas et al., 2014).

The performance of other methods were hindered by their underlying assumptions, such as the RS method, which has the tendency to assign approximately the same proportion of spikes as being within bursts in each spike train, regardless of how spikes are distributed. This meant that the RS method tended to both overestimate bursting activity in non-bursting trains and underestimate bursting in spike trains in which most spikes occurred within bursts, making it unsuitable for analyzing MEA recordings in which the level of bursting activity does not remain consistent across all electrodes.

The CMA method, which was designed for the purpose of analyzing recordings from developing human neuronal networks, was a promising candidate in our analysis. The major limitation of this method was its tendency to erroneously detect a large amount of bursting activity in spike trains containing no or sparse bursting activity, in particular

those with non-stationary firing rates. The authors' suggestion for post hoc screening can address this issue, but also leads to underestimation of bursting in some spike trains, as it does not allow for any shorter bursts to be identified in spike trains in which long erroneous bursts were initially detected by the CMA method.

Based on our analysis, two burst detectors showed the most promise, namely the MI and logISI methods. These methods possessed the majority of properties we deemed desirable for a burst detector and were generally able to achieve high coherence with visually detected bursts in experimental data when their parameter values were chosen optimally. These methods, however, still had limitations; the MI method requires the choice of a large number of parameters, the correct value of which can be challenging to determine, and the logISI method had a tendency to underestimate burst durations in some cases.

Given that we have found no 'perfect' method for burst detection, our advice is to choose a burst detector based on the number of degrees of freedom the user wishes to control. The MI method consistently outperformed throughout our analysis, and is our recommendation for a first choice when selecting a burst detector. Although it has a significant number of parameters to be set by the user, unlike methods with probabilistic thresholds, these parameters are easy to interpret and adjust to achieve the desired burst detection results. If appropriate parameters cannot be found for this method, a high performing alternative is the logISI method, which can be implemented without choosing any input parameters. This method is most effective when there is a clear distinction between the sizes of within-burst and between-burst intervals. In cases when this distinction is not apparent, we recommend the PS and CMA methods as reasonably effective alternatives. Due to their tendency to overestimate burst durations in some circumstances, however, post hoc screening for outliers in terms of burst duration is advisable when using these methods.

The most robust approach to burst detection would be to apply a number of burst detectors to the data of interest, and compare the result of each method using summary

statistics or measures such as the Hamming distance. If the methods are in agreement, this provides confidence in the conclusions about the nature of bursting activity in the experimental data. Any major discrepancies between the methods can also be used to pinpoint areas where one or more burst detectors may be performing poorly, an issue that can be further investigated through visual inspection of the specific spike trains of interest. In particular, periods in which the bursts found by the MI method deviate greatly from those found by other methods may suggest that the MI method parameters used were suboptimal for the analysis of these spike trains. In general, we found that for spike trains that are easy to annotate using visual inspection, high performing burst detectors tend to be in close agreement. However, in spike trains for which two humans may not be able to agree on how to appropriately allocate spikes to bursts, it is likely that the methods will also disagree, and discretion is required.

By employing this method of applying a number of burst detection techniques to recordings of networks of hiPSC-derived neurons over a range of developmental ages, we found that bursting arises in a majority of these networks around eight to ten weeks after differentiation. This is a similar time frame to the findings from some previous studies of human stem cell-derived neuronal networks (Heikkilä et al., 2009; Kirwan et al., 2015). Additionally, although we observed some increase in bursting activity over development, the rate of this increase was far lower than that which has been commonly seen in developing rodent neuronal networks (Chiappalone et al., 2005; Charlesworth et al., 2015; Wagenaar et al., 2006).

One limitation of our study was the limited number of burst detectors examined. This was a deliberate choice, due to the extensive number of burst detectors available in the literature, which makes an exhaustive analysis of all methods impossible. Instead of providing a brief analysis of all burst detection methods, we restricted the scope of our study in order to provide a thorough assessment of what we saw as the most promising methods of burst detection, and to offer implementable recommendations to researchers working in this area. As an accompaniment, we also provide R code to implement all of the meth-

ods examined here.

The results of our study were also influenced by how the 'ground truth' bursts were chosen by visual inspection in the experimental RGC recordings, which is necessarily a subjective choice. However, the relatively high degree of coherence between our visually annotated bursts and those identified by a number of burst detectors suggests that our definition of bursts was largely similar to that of other authors.

There are several avenues through which this work could be extended. One area that we did not explore is the possibility of improving the results of burst detection by using a pre-processing step (Martens et al., 2014). Also, during our analysis, ideas arose about how the methods under review could be improved to enhance their performance. For example, for the CMA method, restricting the allowed values for $maxISI$ to within a biologically realistic range may reduce the method's tendency to overestimate bursting in non-bursting spike trains and remove the need for post hoc screening. However, to ensure a fair and unbiased assessment of different burst detectors, we restricted our study to the original implementation of the authors' methods. Future studies in this area could look at how altering the existing methods could improve their performance.

Another area for consideration relates to which features of bursts are the most informative to extract. In past studies of rodent neuronal networks, we have shown that the temporal structure of bursting activity, measured by the CV of IBI, can be an important feature in distinguishing different types of network activity (Charlesworth et al., 2015). However, in the human network recordings examined here, we found no strong relationship between the CV of IBI and developmental age. A greater understanding of which are the most distinguishing features of bursts in human neuronal networks may inform future approaches to burst detection in these networks.

Acknowledgements

We thank Tokdar et al., Pasquale et al., Ko et al. and Gourévitch et al. for code used in this study and Matthias Hennig, Timothy Shafer, Diana Hall and Catherine Cutts for comments on the manuscript. Present address for CWT: Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, UK.

Grants

Experimental data collection was supported by the BBSRC (PC, OP, grant number BB/H008608/1). EC was supported by a Wellcome Trust PhD Studentship and NIHR Cambridge Biomedical Research Centre Studentship. CWT was supported by a bursary from the Bridgwater Summer Undergraduate Research programme.

Disclosures

The authors declare no competing financial interests.

References

Bakkum DJ, Radivojevic M, Frey U, Franke F, Hierlemann A, Takahashi H. Parameters for burst detection. *Front Comput Neurosci* 7: 193, 2013.

Charlesworth P, Cotterill E, Morton A, Grant SG, Eglen SJ. Quantitative differences in developmental profiles of spontaneous activity in cortical and hippocampal cultures. *Neural Dev* 10: 1–10, 2015.

Charlesworth P, Morton A, Eglen SJ, Komiyama NH, Grant SG. Canalization of genetic and pharmacological perturbations in developing primary neuronal activity patterns. *Neuropharmacol* 100: 47–55, 2016.

Chen L, Deng Y, Luo W, Wang Z, Zeng S. Detection of bursts in neuronal spike trains by the mean inter-spike interval method. *Prog Nat Sci* 19: 229–235, 2009.

Chiappalone M, Novellino A, Vajda I, Vato A, Martinoia S, van Pelt J. Burst detection algorithms for the analysis of spatio-temporal patterns in cortical networks of neurons. *Neurocomputing* 65-66: 653–662, 2005.

Cocatre-Zilgien J, Delcomyn F. Identification of bursts in spike trains. *J Neurosci Methods* 41: 19–30, 1992.

Demas J, Eglen SJ, Wong ROL. Developmental loss of synchronous spontaneous activity in the mouse retina is independent of visual experience. *J Neurosci* 23: 2851–2860, 2003.

Eglen S, Weeks M, Jessop M, Simonotto J. A data repository and analysis framework for spontaneous neural activity recordings in developing retina. *Gigascience* 3:3, 2014.

Eisenman LN, Emnett CM, Mohan J, Zorumski CF, Mennerick S. Quantification of bursting and synchrony in cultured hippocampal neurons. *J Neurophysiol* 114: 1059–1071, 2015.

Gilchrist KH, Lewis GF, Gay EA, Sellgren KL, Grego S. High-throughput cardiac safety evaluation and multi-parameter arrhythmia profiling of cardiomyocytes using microelectrode arrays. *Toxicol Appl Pharmacol* 288: 249–257, 2015.

Gourévitch B, Eggermont JJ. A nonparametric approach for detection of bursts in spike trains. *J Neurosci Methods* 160: 349–58, 2007.

Heikkilä TJ, Ylä-Outinen L, Tanskanen JMA, Lappalainen RS, Skottman H, Suuronen R, Mikkonen JE, Hyttinen JAK, Narkilahti S. Human embryonic stem cell-derived neuronal cells form spontaneously active neuronal networks in vitro. *Exp Neurol* 218: 109–16, 2009.

Hennig MH, Grady J, van Coppenhagen J, Sernagor E. Age-dependent homeostatic plasticity of GABAergic signaling in developing retinal networks. *J Neurosci* 31: 12159–64, 2011.

Kaneoke Y, Vitek J. Burst and oscillation as disparate neuronal properties. *J Neurosci Methods* 68: 211–223, 1996.

Kapucu FE, Tanskanen JMA, Mikkonen JE, Ylä-Outinen L, Narkilahti S, Hyttinen JAK. Burst analysis tool for developing neuronal networks exhibiting highly varying action potential dynamics. *Front Comput Neurosci* 6: 38, 2012.

Kirwan P, Turner-Bridger B, Peter M, Momoh A, Arambepola D, Robinson HPC, Livesey FJ. Development and function of human cerebral cortex neural networks from pluripotent stem cells in vitro. *Dev* 142: 3178–3187, 2015.

Ko D, Wilson CJ, Lobb CJ, Paladini CA. Detection of bursts and pauses in spike trains. *J Neurosci Methods* 211: 145–58, 2012.

Legéndy CR, Salcman M. Bursts and recurrences of bursts in the spike trains of spontaneously active striate cortex neurons. *J Neurophysiol* 53: 926–39, 1985.

Lisman JE. Bursts as a unit of neural information: Making unreliable synapses reliable. *Trends Neurosci* 20: 38–43, 1997.

Maccione A, Hennig MH, Gandolfo M, Muthmann O, van Coppenhagen J, Eglen SJ, Berdondini L, Sernagor E. Following the ontogeny of retinal waves: pan-retinal recordings of population dynamics in the neonatal mouse. *J Physiol* 592: 1545–63, 2014.

Maeda E, Robinson HP, Kawana A. The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. *J Neurosci* 15: 6834–6845, 1995.

Martens MB, Chiappalone M, Schubert D, Tiesinga PHE. Separating burst from background spikes in multichannel neuronal recordings using return map analysis. *Int J Neural Syst* 24: 1450012, 2014.

Mazzoni A, Broccard FD, Garcia-Perez E, Bonifazi P, Ruaro ME, Torre V. On the dynamics of the spontaneous activity in neuronal networks. *PLoS One* 2: e439, 2007.

Mukai Y, Shiina T, Jimbo Y. Continuous monitoring of developmental activity changes in cultured cortical networks. *Electr Eng Jpn* 145: 28–37, 2003.

Nex Technologies 2014 Neuroexplorer manual (`http://www.neuroexplorer.com/downloads/NeuroExplorerManual.pdf`).

Nicolas J, Hendriksen PJM, van Kleef RGDM, de Groot A, Bovee TFH, Rietjens IMCM, Westerink RHS. Detection of marine neurotoxins in food safety testing using a multielectrode array. *Mol Nutr Food Res* 58: 2369–2378, 2014.

Odawara A, Katoh H, Matsuda N, Suzuki I. Induction of long-term potentiation and depression phenomena in human induced pluripotent stem cell-derived cortical neurons. *Biochem Biophys Res Commun* 469: 856 – 862, 2016.

Pasquale V, Martinoia S, Chiappalone M. A self-adapting approach for the detection of bursts and network bursts in neuronal cultures. *J Comput Neurosci* 29: 213–229, 2010.

Pimashkin A, Kastalskiy I, Simonov A, Koryagina E, Mukhina I, Kazantsev V. Spiking signatures of spontaneous activity bursts in hippocampal cultures. *Front Comput Neurosci* 5: 1–12, 2011.

Potter SM, DeMarse TB. A new approach to neural cell culture for long-term studies. *J Neurosci Methods* 110: 17–24, 2001.

R Core Team 2015 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

Selinger JV, Kulagina NV, O'Shaughnessy TJ, Ma W, Pancrazio JJ. Methods for characterizing interspike intervals and identifying bursts in neuronal activity. *J Neurosci Methods* 162: 64–71, 2007.

Shi Y, Kirwan P, Livesey FJ. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat Protoc* 7: 1836–1846, 2012.

Tam D. An alternate burst analysis for detecting intra-burst firings based on inter-burst periods. *Neurocomputing* 46: 1155–1159, 2002.

Tokdar S, Xi P, Kelly RC, Kass RE. Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *J Comput Neurosci* 29: 203–12, 2010.

Turnbull L, Dian E, Gross G. The string method of burst identification in neuronal spike trains. *J Neurosci Methods* 145: 23–35, 2005.

Valdivia P, Martin M, LeFew WR, Ross J, Houck KA, Shafer TJ. Multi-well microelectrode array recordings detect neuroactivity of ToxCast compounds. *Neurotoxicology* 44: 204–217, 2014.

van Elburg RA, van Ooyen A. A new measure for bursting. *Neurocomputing* 58-60: 497–502, 2004.

Wagenaar D, DeMarse T, Potter S 2005 Meabench: A toolset for multi-electrode data acquisition and on-line analysis In: *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005*, p. 518–521.

Wagenaar DA, Pine J, Potter SM. An extremely rich repertoire of bursting patterns during the development of cortical cultures. *BMC Neurosci* 7: 11, 2006.

Weihberger O, Okujeni S, Mikkonen JE, Egert U. Quantitative examination of stimulus-response relations in cortical networks in vitro. *J Neurophysiol* 109: 1764–1774, 2013.

Weyand TG, Boudreaux M, Guido W. Burst and tonic response modes in thalamic neurons during sleep and wakefulness. *J Neurophysiol* 85: 1107–1118, 2001.

Xia Y, Gopal KV, Gross GW. Differential acute effects of fluoxetine on frontal and auditory cortex networks in vitro. *Brain Res* 973: 151–160, 2003.

List of Figures

5 ROC curves showing the fraction of true positive (sensitivity) and false positive spikes (1-specificity) identified as being within bursts for a variety of input parameter values, for recordings of mouse retinal ganglion cells at **A** P9 and **B** P15. The ground truth bursts for hour-long recordings from five randomly selected electrodes at each age were determined by visual inspection (examples in Figure 2), and the mean performance of each burst detector over the five electrodes is shown. Some curves do not span the entire range because of innate restrictions on the maximum proportion of spikes which can be allocated to bursts by each method. The green dot represents the single specificity and sensitivity value found by the CMA method, which has no obvious parameter to vary.

6 Detailed analysis of mouse retinal ganglion cell recordings. **A** Fraction of spikes in bursts, and **B** Mean burst duration found by each burst detector. Each electrode was counted as one data point in the box plots. The legend in A applies to both A and B. **C** Bursts detected by each burst detector over a 120 s sample of a P15 spike train, and **D** 15 s sample showing the first bursting episode from the same spike train. Horizontal bars in C and D denote the bursts detected by each method. Blue bars above the spike train represent the bursts annotated by a human observer.

7 Analysis of recordings of networks of human induced pluripotent stem cell-derived neurons. **A** Mean burst duration, **B** Fraction of spikes in bursts, and **C** Coefficient of variation of interburst intervals (CV of IBI). Each data point in the box plots is the mean value across all electrodes from one recording. **D** Median normalized Hamming distance between each pairwise combination of burst detection methods at each week after plating.

8 Results of the four burst detectors applied to samples of human induced pluripotent stem cell-derived neuronal network recordings at **A, B** 4 weeks after plating (WAP), **C, D** 8 WAP, and **E, F** 12 WAP. Spike trains on the left show 30 s of activity, with the scale bar representing 3 s. The inset on the right of each spike train is an enlarged version of the last 3 s of this activity. Horizontal bars denote the bursts detected by each method.

List of Tables

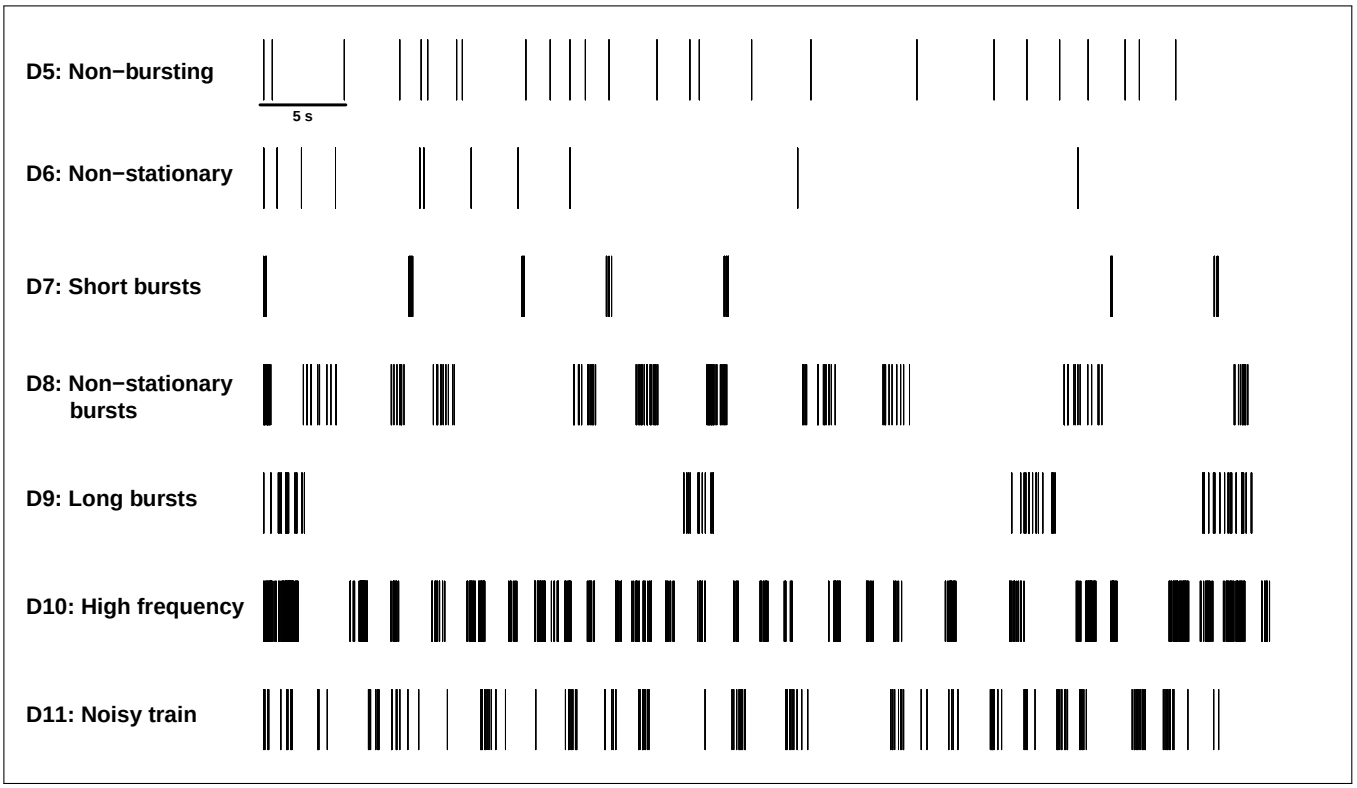

Figure 1: One-minute examples of simulated spike trains for evaluating desirable features D5–D11. Scale bar represents 5 s.

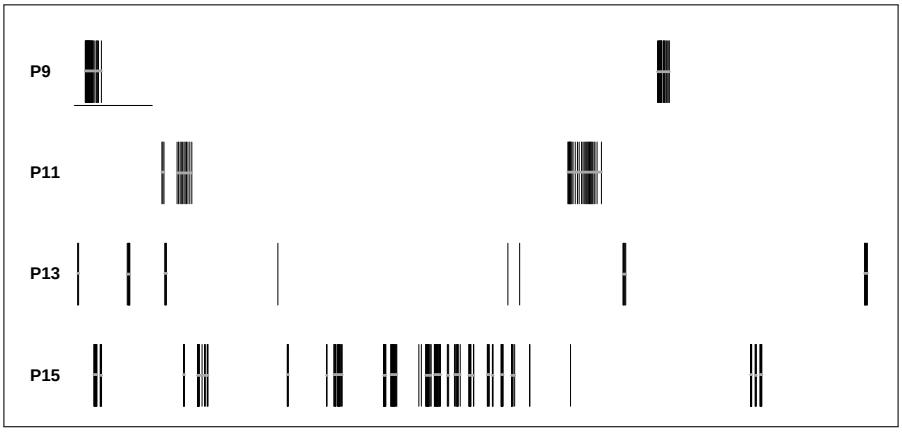

Figure 2: Examples of one-minute spike trains from recordings of mouse retinal ganglion cell at each postnatal day. Horizontal bars represent bursts annotated by a human observer. Scale bar represents 5 s of activity.

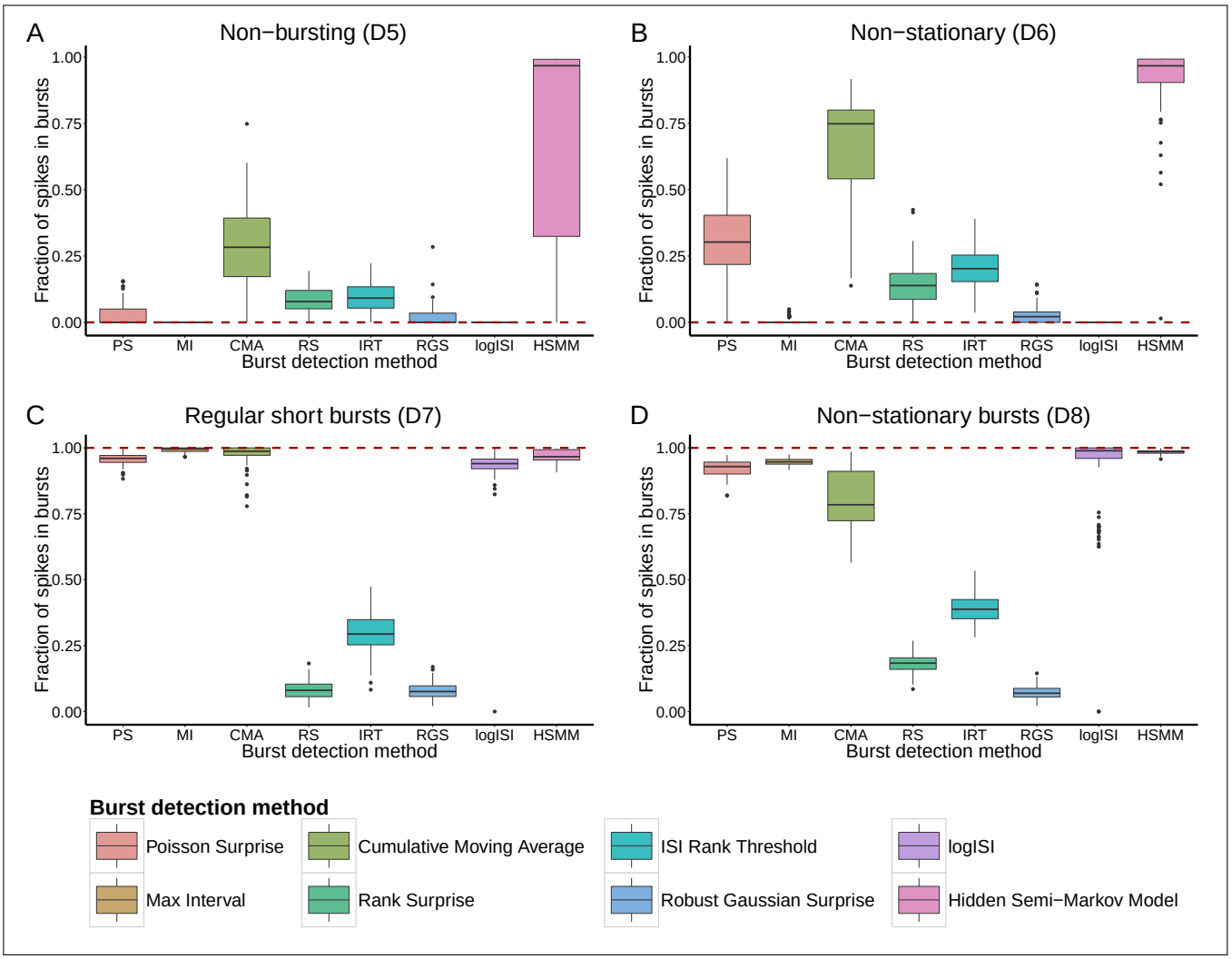

Figure 3: Fraction of spikes in bursts found by each burst detector in 100 synthetic trains with **A** No bursting (D5), **B** No bursting and non-stationary firing rate (D6), **C** Short regular bursts (D7), **D** Bursts with non-stationary burst lengths and durations (D8). Dotted line shows desired result from an ideal burst detector; methods close to this line are deemed to work well. In each 'box-and-whisker' plot, boxes show the median $\pm$ inter-quartile range (IQR), and whiskers extend to median $\pm$ 1.5$\times$ IQR. Outliers are represented as points.

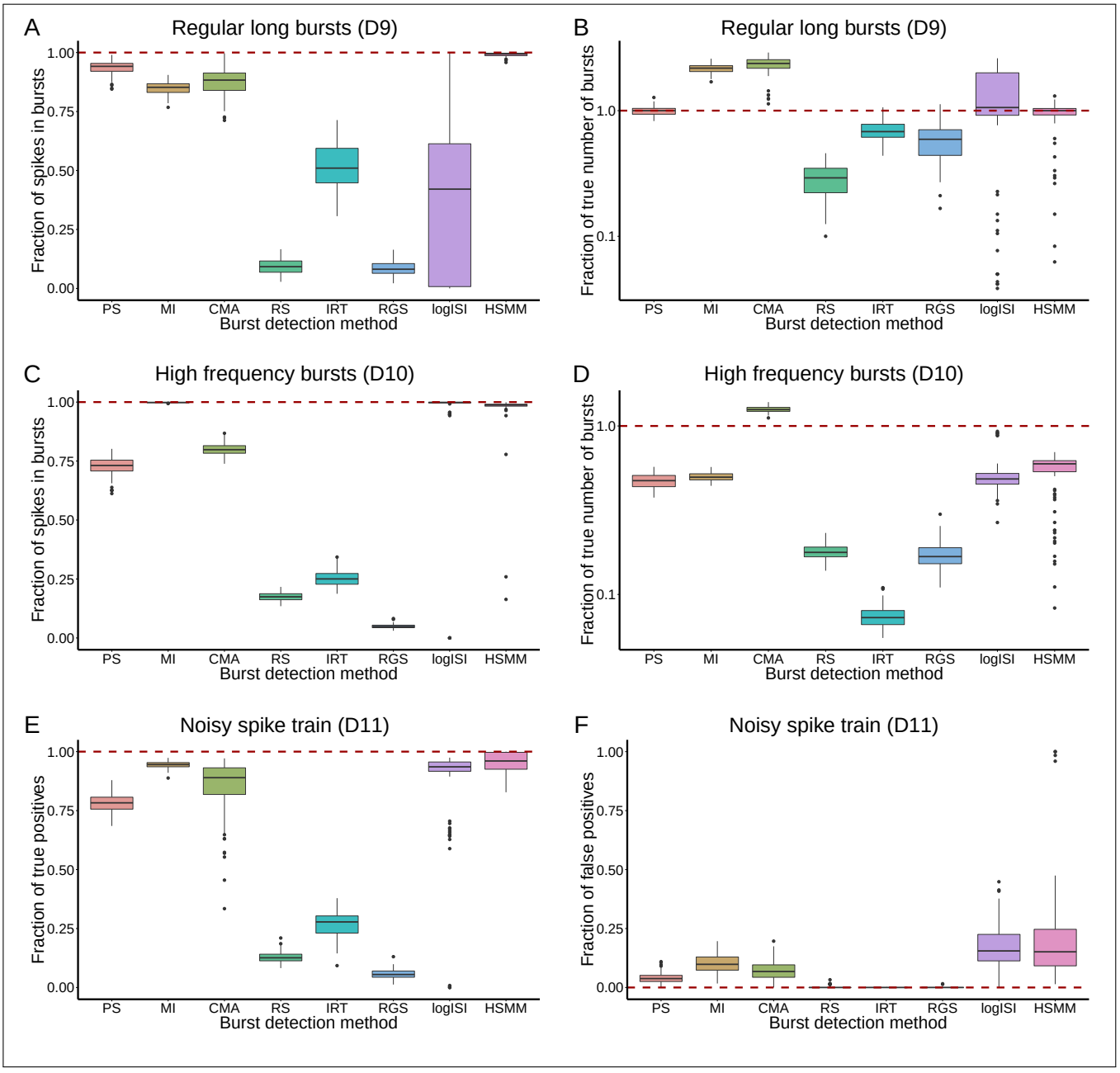

Figure 4: Results of each burst detector at analyzing 100 synthetic spike trains. **A** Fraction of spikes in bursts, and **B** Fraction of true number of bursts in spike trains with regular long bursts (D9); **C** Fraction of spikes in bursts, and **D** Fraction of true number of bursts in spike trains with high frequency bursting (D10); **E** Fraction of true positive, and **F** Fraction of false positive spikes in bursts in spike trains containing both bursting and noise (D11). Values calculated as outlined in the methods. Box plots and dotted line as per Figure 3 legend. **B** and **D** are presented on a log-scale.

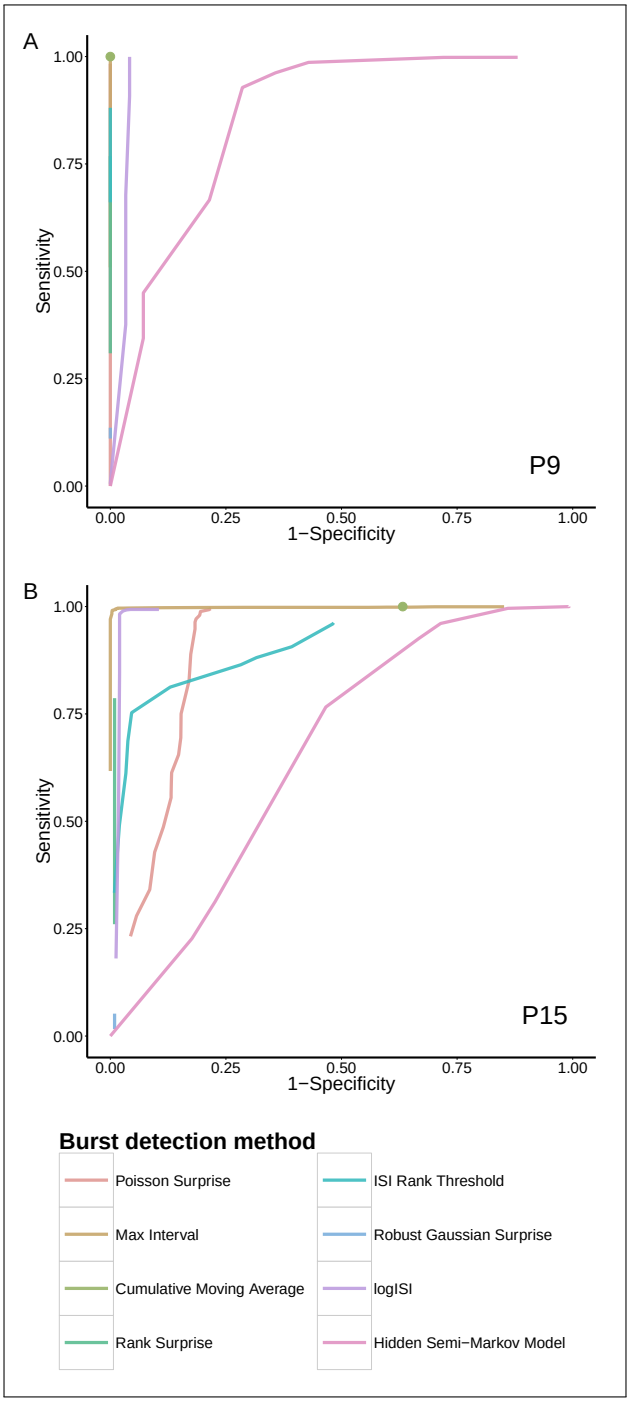

Figure 5: ROC curves showing the fraction of true positive (sensitivity) and false positive spikes (1-specificity) identified as being within bursts for a variety of input parameter values, for recordings of mouse retinal ganglion cells at **A** P9 and **B** P15. The ground truth bursts for hour-long recordings from five randomly selected electrodes at each age were determined by visual inspection (examples in Figure 2), and the mean performance of each burst detector over the five electrodes is shown. Some curves do not span the entire range because of innate restrictions on the maximum proportion of spikes which can be allocated to bursts by each method. The green dot represents the single specificity and sensitivity value found by the CMA method, which has no obvious parameter to vary.

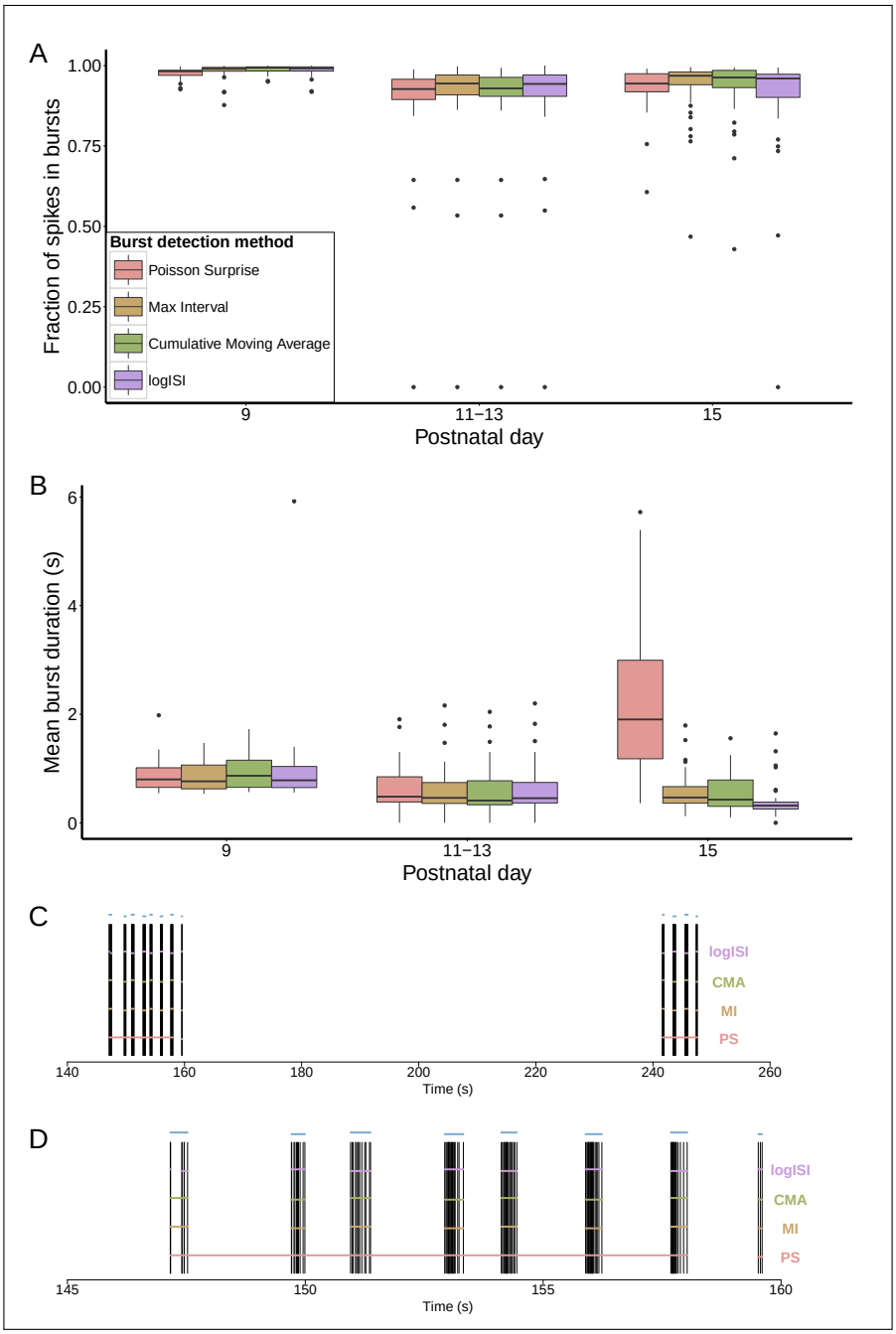

Figure 6: Detailed analysis of mouse retinal ganglion cell recordings. **A** Fraction of spikes in bursts, and **B** Mean burst duration found by each burst detector. Each electrode was counted as one data point in the box plots. The legend in A applies to both A and B. **C** Bursts detected by each burst detector over a 120 s sample of a P15 spike train, and **D** 15 s sample showing the first bursting episode from the same spike train. Horizontal bars in C and D denote the bursts detected by each method. Blue bars above the spike train represent the bursts annotated by a human observer.

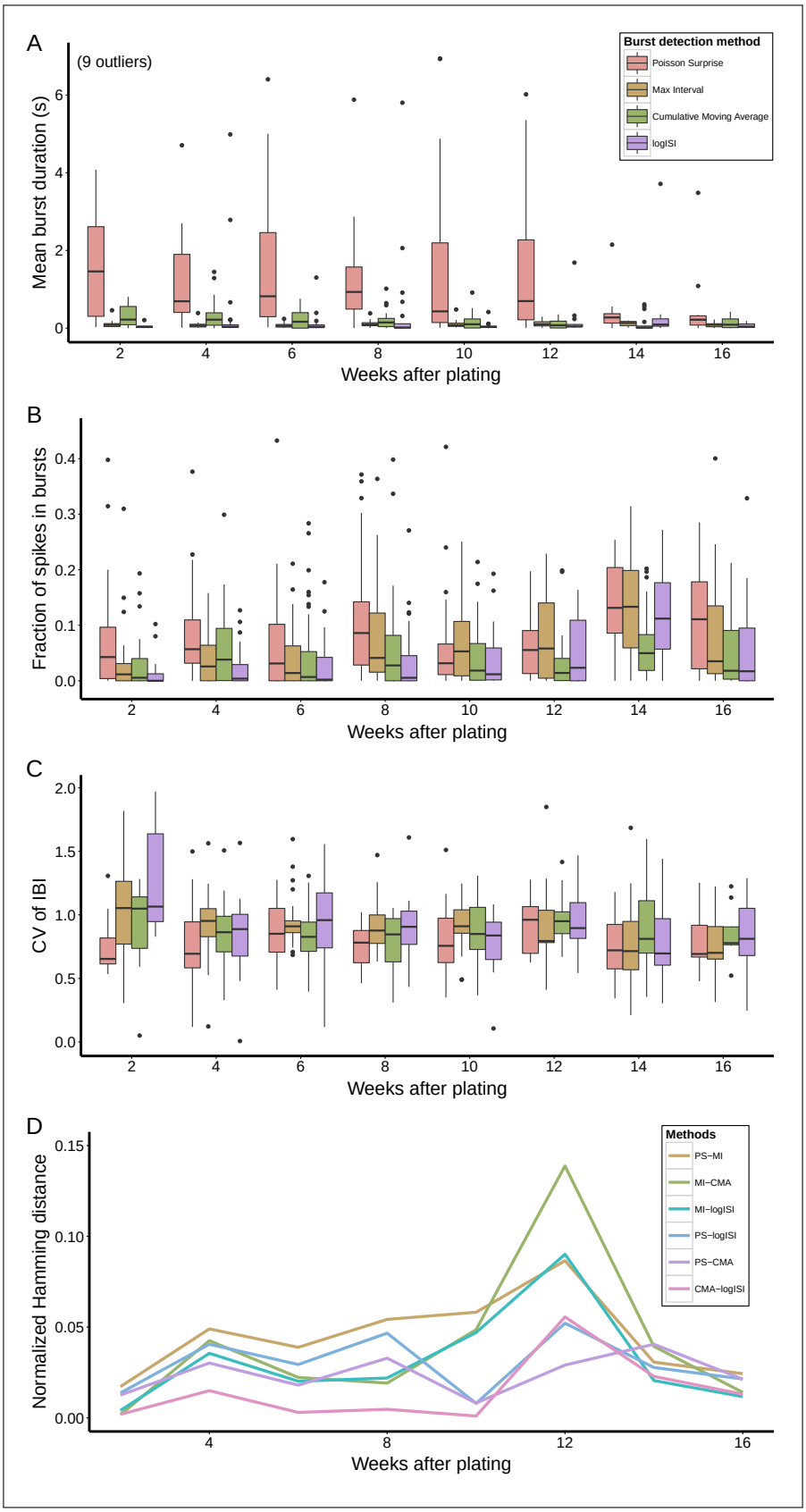

Figure 7: Analysis of recordings of networks of human induced pluripotent stem cell-derived neurons. **A** Mean burst duration, **B** Fraction of spikes in bursts, and **C** Coefficient of variation of interburst intervals (CV of IBI). Each data point in the box plots is the mean value across all electrodes from one recording. **D** Median normalized Hamming distance between each pairwise combination of burst detection methods at each week after plating.

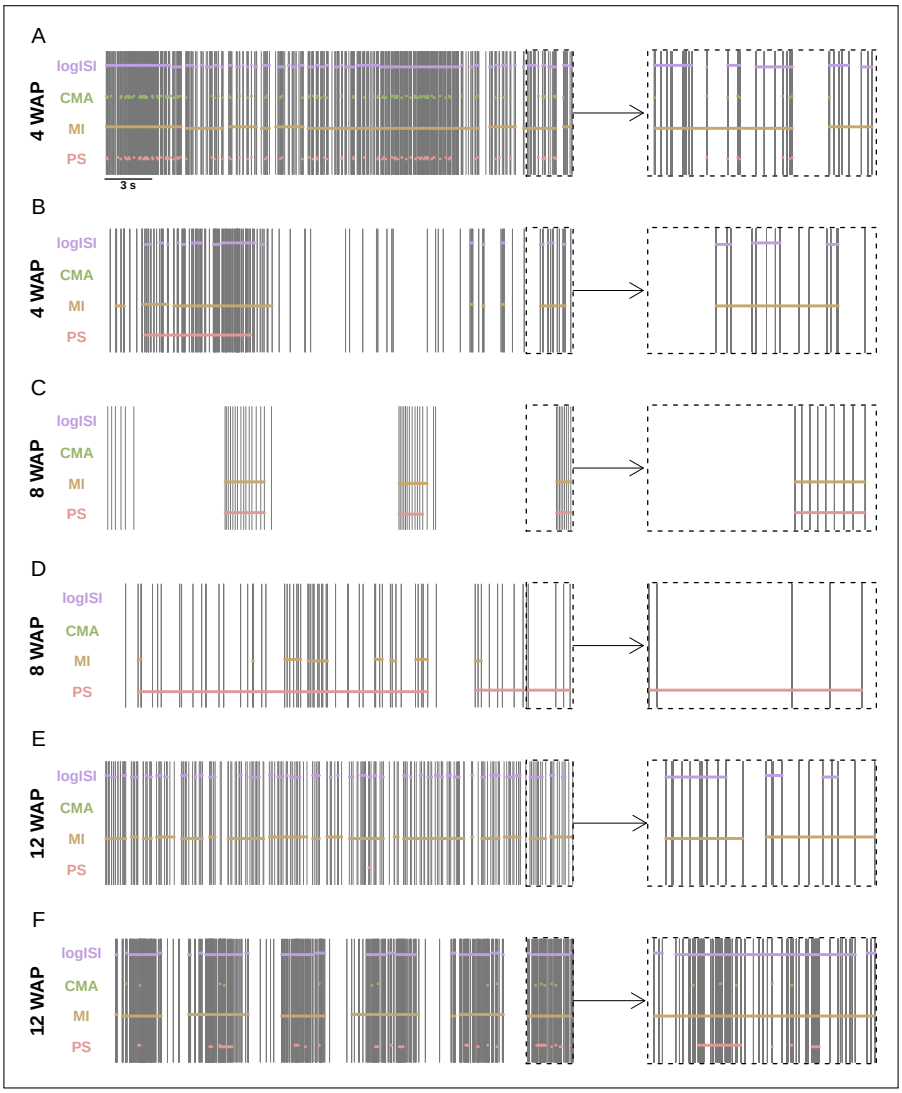

Figure 8: Results of the four burst detectors applied to samples of human induced pluripotent stem cell-derived neuronal network recordings at **A, B** 4 weeks after plating (WAP), **C, D** 8 WAP, and **E, F** 12 WAP. Spike trains on the left show 30 s of activity, with the scale bar representing 3 s. The inset on the right of each spike train is an enlarged version of the last 3 s of this activity. Horizontal bars denote the bursts detected by each method.

Method	Parameter	Value
Poisson Surprise (Legéndy and Salcman, 1985)	Minimum surprise value	$-\log(0.01) \approx 4.6$
MaxInterval (Nex Technologies, 2014)	Maximum beginning ISI Maximum end ISI Minimum interburst interval Minimum burst duration Minimum spikes in a burst	0.17 s 0.3 s 0.2 s 0.01 s 3
Cumulative Moving Average (Kapucu et al., 2012)	$\alpha_1(\alpha_2)$	1.0 (0.5) if skew < 1 0.7 (0.5) if $1 \leq$ skew < 4 0.5 (0.9) if $4 \leq$ skew < 9 0.3 (0.1) if $9 \leq$ skew
Rank Surprise (Gourévitch and Eggermont, 2007)	Largest allowed ISI in burst Minimum surprise value	75th percentile of ISIs $-\log(0.01) \approx 4.6$
ISI Rank Threshold (Hennig et al., 2011)	Rank threshold, θ_R Spike count cutoff, θ_C	0.5 C such that $P(C) = 0.05$
Robust Gaussian Surprise (Ko et al., 2012)	Minimum burst surprise	$-\log(0.01) \approx 4.6$
LogISI (Pasquale et al., 2010)	Maximum cutoff value	100 ms
Hidden Semi-Markov Model (Tokdar et al., 2010)	Probabilistic cutoff Other parameters (N=23)	0.5 As per paper*

Table 1: The eight burst detectors and the parameter values used for the implementation of each method on synthetic spike trains. *These parameters were left set to the default values provided in the 'burstHSMM' R package.

Desirable properties	
D1	Deterministic: The method should detect the same bursts over repeated runs on the same data, to ensure consistency and reproducibility of results
D2	No assumption of spike train distribution: The method should not assume ISIs follow a standard statistical distribution, to ensure wide applicability to a variety of spike trains
D3	Number of parameters: The method should have few parameters, to reduce the variability inherently introduced through parameter choice
D4	Computational time: The method should run in a reasonable amount of time using standard personal computers
D5	Non-bursting trains: The method should detect few spikes as being within bursts in spike trains containing no obvious bursting behaviour
D6	Non-stationary trains: The method should detect few spikes as being within bursts in spike trains with non-stationary firing rates that contain no obvious bursting behaviour
D7	Regular short bursts: The method should detect a high proportion of spikes in bursts in spike trains containing short well-separated bursts
D8	Non-stationary bursts: The method should detect a high proportion of spikes in bursts in spike trains containing bursts with variable durations and numbers of spikes per burst
D9	Regular long bursts: The method should detect a high proportion of spikes in bursts and accurate number of bursts in spike trains containing long bursts with low within-burst firing rates
D10	High frequency bursts: The method should detect a high proportion of spikes in bursts and accurate number of bursts in spike trains containing a large number of short bursts
D11	Noisy train: The method should classify a high number of within-burst spikes as bursting and a low number of interburst spikes as bursting in spike trains containing both bursts and noise spikes

Table 2: Desirable properties for a burst detector.

Spiking model	Property	Parameters	Mean % spikes in bursts
100 Poisson spiking	Computational time (D4)	$\lambda = 1\,\text{Hz}$	0
50 Poisson spiking 50 Gamma distributed ISIs	Non-bursting (D5)	$\lambda = 0.5\,\text{Hz}$, $N = 50$ $\alpha = 1$, $\beta = 0.5$, $N = 50$	0
100 Inhomogeneous Poisson	Non-stationary (D6)	$\lambda(t) = 1 + \frac{1}{300}t$	0
100 Poisson bursting	Short bursts (D7)	$\lambda = 0.2\,\text{Hz}$, $n = 5, r = 0.3\,\text{s}$	100
100 Poisson bursting	Non-stationary bursts (D8)	$\lambda = 0.3\,\text{Hz}$, $n \sim \mathcal{U}(5, 18)$, $r \sim \mathcal{U}(0.3, 3)\,\text{s}$	100
100 Poisson bursting	Long bursts (D9)	$\lambda = 0.1\,\text{Hz}$, $n = 18, r = 3\,\text{s}$	100
100 Poisson bursting	High frequency (D10)	$\lambda = 1\,\text{Hz}$, $n = 10, r = 0.5\,\text{s}$	100
100 Poisson bursting with Gamma distributed noise ISIs	Noisy train (D11)	$\lambda = 0.5\,\text{Hz}$, $n = 8, r = 0.8\,\text{s}$ $\alpha = 1$, $\beta = 0.5$	91

Table 3: Models and parameter values used to generate synthetic spike trains for each desirable property. Each spike train was 300 s duration, and the number, N, of simulated trains was 100, unless otherwise stated. α and β represent the shape and inverse scale parameters of the Gamma distribution, respectively.

	PS	MI	CMA	RS	IRT	RGS	logISI	HSMM
D1 Deterministic	✓	✓	✓	✓	✓	✓	✓	×
D2 Distribution assumption	×	✓	✓	✓	✓	×	✓	×
D3 Number of parameters	✓	×	✓	✓	✓	✓	✓	×
D4 Computational time	✓	✓	✓	✓	✓	✓	✓	×

Table 4: Summary of the performance of each method on desirable properties D1–D4.

	PS	MI	CMA	RS	IRT	RGS	logISI	HSMM
D5 Non-bursting	4	1	7	5	6	3	1	8
D6 Non-stationary	6	2	7	4	5	3	1	8
D7 Regular bursting	4	1	2	7	6	7	5	3
D8 Non-stationary bursts	4	3	5	7	6	8	2	1
D9 Long bursts	2	4	3	8	5	7	6	1
D10 High frequency	5	1	4	7	6	8	2	3
D11 Noisy bursts	5	1	2	7	6	8	4	2
Total (Relative rank)	30 (4)	13 (1)	30 (4)	45 (8)	40 (6)	44 (7)	21 (2)	26 (3)

Table 5: The relative rank of the performance of each method on desirable properties D5–D11 (1=best, 8=worst).

Method	Parameter	Value
Poisson Surprise	Minimum surprise value	$-\log(0.0025) \approx 6$
MaxInterval	Maximum beginning ISI	0.2 s
	Maximum end ISI	0.3 s
	Minimum interburst interval	0.2 s
	Minimum burst duration	0.01 s
	Minimum spikes in a burst	3
Cumulative Moving Average	α_1	1.0 if skew $<$ 1
		0.7 if $1 \leq$ skew $<$ 4
		0.5 if $4 \leq$ skew $<$ 9
		0.3 if $9 \leq$ skew
	Maximum mean burst duration	5 s
	Maximum mean spikes per burst	50
LogISI	Maximum cutoff value	150 ms

Table 6: Parameter values used for burst detection on human induced pluripotent stem cell-derived neuronal networks.