Unsupervised spike sorting for large scale, high density multielectrode arrays

Gerrit Hilgen[1,*], Martino Sorbaro[2,3,*], Sahar Pirmoradian[2], Jens-Oliver Muthmann[2,4,5], Ibolya E. Kepiro[6,7], Simona Ullo[8], Cesar Juarez Ramirez[2], Alessandro Maccione[9], Luca Berdondini[9], Vittorio Murino[8], Diego Sona[8], Francesca Cella Zanacchi[6], Upinder S. Bhalla[5], Evelyne Sernagor[1], Matthias H. Hennig[2]

[*]equal contribution
[1]Institute of Neuroscience, Newcastle University, Newcastle, UK
[2]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, UK
[3]Department of Computational Biology, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden
[4]Manipal University, Manipal, India
[5]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India
[6]Nanophysics (NAPH), Istituto Italiano di Tecnologia, Genova, Italy
[7]Faculty of Science Engineering and Computing, Kingston University, London, UK
[8]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
[9]Neuroscience and Brain Technologies (NBT), Istituto Italiano di Tecnologia, Genova, Italy

Correspondence: Matthias H. Hennig, m.hennig@ed.ac.uk

Abstract

A new method for automated spike sorting for recordings with high density, large scale multielectrode arrays is presented. It is based on an efficient, low-dimensional representation of detected events by their estimated spatial current source locations and dominant spike shape features. Millions of events can be sorted in just minutes, and the full analysis chain scales roughly linearly with recording time. We demonstrate this method using recordings from the mouse retina with a 4,096 channel array, and present validation based on anatomical imaging and model-based quality control. Our analysis shows that it is feasible to reliably isolate the activity of hundreds to thousands of retinal ganglion cells in single recordings.

Introduction

Large scale, dense probes and arrays and planar multielectrode arrays now make it possible to perform extracellular recordings from potentially thousands of neurons simultaneously (Eversmann et al., 2003; Berdondini et al., 2005; Hutzler et al., 2006; Frey et al., 2010; Ballini et al., 2014; Maccione et al., 2014; Müller et al., 2015; Obien et al., 2015). Obtaining and analyzing such data has many advantages. For example, only large unbiased samples of neural activity from single preparations enable an unbiased characterization of neural variability, and they can reliably distinguish between variability within and between preparations. Moreover, for the first time it is possible to systematically examine hypotheses and models of distributed encoding and representations.

These applications require reliable isolation of extracellularly recorded spikes generated by single neurons, a process called spike sorting (reviewed by Rey et al., 2015). For conventional devices with tens of recording channels, a typical workflow consists of an initial event detection step, followed by semi-automated spike sorting and manual inspection and refinement of the proposed event assignments to single units. If the recording channels are sufficiently well separated, there will be no or little overlap between their signals, and spike sorting can be performed by clustering a low-dimensional representation of spike shapes, which differ for individual neurons according to their relative location to the electrode (e.g. Lewicki, 1998; Harris et al., 2000; Quiroga et al., 2004).

This approach, however, becomes unfeasible for dense, large scale recordings for two rather obvious reasons. First, the sheer size of the data sets makes extensive manual intervention impractical, hence as much of the process as possible, including quality control, should be automated. Second, on dense arrays spike sorting becomes a very complex assignment problem, since not only multiple neurons contribute to the compound signal recorded on individual channels, but each neuron's spikes may also be recorded by several neighboring channels simultaneously (Prentice et al., 2011; Rossant et al., 2016). Therefore, each detected spike is a high dimensional event, which ideally should be sorted using the full spatio-temporal footprint it leaves on multiple channels, a computationally challenging task. Existing promising solutions that could potentially be scaled up to thousands of channels are template matching methods (Prentice et al., 2011; Marre et al., 2012), and a recently developed method that shrinks the feature space such that fitting of a mixture model becomes feasible for larger data sets (Rossant et al., 2016).

Here we present a new, very fast and fully automated method for spike sorting that relies on a highly efficient representation of the relevant information contained in the raw signal. The dense sampling enables us, in a first step, to use the spatial signal decay to obtain a rough estimate of the current source location for each detected event (Muthmann et al., 2015). Events detected this way tend to form dense and spatially well-separated clusters, which, as we show using optogenetic stimulation and confocal imaging, are spikes that usually originate from single neurons. In addition, for each event an average waveform is obtained, with noise reduced by signal interpolation. Features extracted from this single waveform are then combined with the spatial location, such that the clustering problem is reduced to finding local density peaks in few dimensions. This is achieved by parallelized clustering algorithm, which is capable of sorting millions of spikes within a few minutes on a fast workstation.

We tested this method on recordings from the mouse retina with a 4,096 channel array. To evaluate the quality of the sorting, post hoc analysis is performed on the clustered data. This analysis allows largely unsupervised rejection of poorly clustered units, and highlights potentially problematic cases for further inspection. The software, including a tool for quick visualization of the sorted data, can be downloaded at `https://github.com/martinosorb/herding-spikes`.

Methods

Electrophysiology

Experimental procedures were approved by the UK Home Office, Animals (Scientific procedures) Act 1986 and performed at Newcastle University, UK. C3H/HeNHsd mice (also known as rd1 mice) purchased from Harlan Laboratories (Indianapolis, USA) were crossbred with B6.Cg-Tg(Thy1-COP4/EYFP)9Gfng/J (ChR2) mice, purchased from the Jackson Laboratory (Bar Harbor, USA) (for details see Barrett et al., 2015). Experiments were performed on adult C57bl/6 (P27-39) and ChR2 mice (P69-P96). High density recordings from the retinal ganglion cell (RGC) layer were performed using the BioCam4096 platform with APS MEA chips type BioChip 4096S (3Brain GmbH, Switzerland), providing 4096 square microelectrodes (21 µm x 21 µm) on an active area of 2.67 mm x 2.67 mm, aligned in a square grid with 42 µm spacing. The platform records at a sampling rate of about 7 kHz/electrode when measuring from the full 64 x 64 electrode array, and at 24 kHz when recording from one quarter of all electrodes. Raw data were visualized and recorded with the BrainWave software provided with the BioCam4096 platform. Activity was recorded at 12 bits resolution per electrode, low-pass filtered at 5 kHz with the on-chip filter and high-pass filtered by setting the digital high-pass filter of the platform at 0.1 Hz.

Mice were killed by cervical dislocation and enucleated prior to retinal isolation. The isolated retina was placed, RGC layer facing down, onto the MEA. Coupling between the tissue and the electrodes was achieved by flattening the retina on the array under a small piece of polyester membrane filter (Sterlitech, Kent, WA) maintained in place by a stainless steel ring. The retina was kept at 32 °C with an in-line heather (Warner Instruments) and continuously perfused using a peristaltic pump (∼1 ml/min) with artificial cerebrospinal fluid (aCSF) containing the following (in mM): 118 NaCl, 25 $NaHCO_3$, 1 NaH_2 PO_4, 3 KCl, 1 $MgCl_2$, 2 $CaCl_2$, and 10 glucose, equilibrated with 95 % O_2 and 5 % CO_2. All preparations were performed under dim red light and the room was maintained in darkness throughout the experiment.

Visual stimulation

Visual stimuli (664x664 pixel images for a total area of 2.67x2.67 mm) were presented using a custom built high-resolution photostimulation system based on a DLP video projector (lightCrafter, Texas Instruments, USA) combined with a custom made photostimulation software and synchronized with the recording system. Neutral density filters (4.5 - 1.9) were used to control the amount of light falling on the retina.

Photoreceptor-driven responses were acquired at a maximum irradiance of 4 μW/cm^2 (ND 4.5), low enough to avoid eliciting ChR2-driven responses in the ChR2 retinas. To isolate ChR2 responses from photoreceptor-driven responses in these same retinas, we decreased synaptic transmission by increasing the MgCl2 concentration to 2.5mM and by decreasing the $CaCl_2$ concentration to 0.5mM in the aCSF solution, and used 20 μm DNQX, and 20 μm L-AP4 (Tocris Bioscience, UK) to respectively block metabotropic and ionotropic glutamate receptors. We used the broad RGB spectrum of the DLP projector with a maximum irradiance of 0.87 mW/cm^2 (ND 2.2) to evoke ChR2 responses. The stimuli were simple full field flashes of 2 s duration for both the bright and dark phase.

Spike detection, localization and selection

The procedures for spike detection and current source localization were described in detail elsewhere (Muthmann et al., 2015). Briefly, first weighted interpolated signals were generated using two spatial templates to capture both spikes originating either close to or between electrodes. A five channel template with a strong relative weight for the central channel and weaker weights for the four surrounding channels emphasized current sources close to electrodes. Sources between electrodes were captured by a four channel template. A running estimate of the signal baseline noise level was computed from percentiles for signals filtered by both templates for each location, and putative spikes were detected as threshold crossings. Next, the current source location for each event was estimated by computing a weighted center of mass of a baseline-subtracted and thresholded signal. This estimate had a small bias towards electrodes for current sources located between channels, and was less precise for small event amplitudes as the effect of noise became more dominant (Muthmann et al., 2015). Yet spikes showed clear spatial clustering, which enabled the spike sorting based on current source location described below.

Detection was performed with a low threshold to reduce false negatives. This also returned events that were clearly not spikes, but nevertheless crossed threshold due to noise. We found that additional shape criteria to remove such events during the detection phase increased the fraction of false negatives, as it was difficult to determine these parameters a priori. Therefore, we implemented an automated post hoc rejection of events, which was based on a classifier trained on examples of noise events and true spikes from the same data set.

To this end, noise events were sampled from areas on the MEA where no activity was recorded, for instance where incisions were made and no tissue covered the MEA. Generally, such areas could be identified by a very low spike count. Here this was implemented by computing the spike count in 64x64 spatial bins, and taking up to 1000 events from the lowest 0.5 percentile of these bins as examples of noise. A further 1000 events with large amplitudes (>3.5) were used as examples of true spikes. The labeled projections along the main (usually four) principal components, computed from all events, were then taken to train a Support Vector Machine with radial basis functions. This model was finally used to classify events as true spikes or noise.

Spike clustering

Data points were clustered together using an implementation of the mean shift algorithm (Comaniciu and Meer, 2002) available in the scikit-learn open source machine learning library (Pedregosa et al., 2011). Importantly, this algorithm did not require the knowledge of the desired number of clusters; it depended, instead, on a single parameter, the bandwidth h, which determined the expected cluster size. To combine spatial and waveform information, the clustering process was run on a four-dimensional space consisting of two dimensions indicating the location of each event on the chip, x and y, and two dimensions representing the first two principal components of the event's waveform. The latter were multiplied by an additional dimensional constant α that tuned the relative importance of the waveform components compared to the spatial coordinates. We developed a parallelized implementation of this algorithm where multiple data points were simultaneously clustered, which improved performance roughly proportionally to the number of available CPUs.

Quality metric

Following Hill et al. (2011), we fitted a multivariate Gaussian mixture model to a set of N clusters, then estimated cluster overlap using posterior probabilities to obtain the probability of incorrect assignments under the assumption of a Gaussian cluster shape.

Let the probability that spike s is in cluster c be $P(C = c|S = s)$: the estimated fraction of spikes in cluster k that could belong to cluster i is given by $f^p(k, i) = \frac{1}{N_k}\sum_{s\in k} P(C = i|S = s)$; by generalizing to all other clusters we obtained the fraction of false positives in k:

$$\begin{aligned} f^p_k &= \sum_{i\neq k} f^p(k;i) \\ &= \frac{1}{N_k}\sum_{i\neq k}\sum_{s\in k} P(C = i|S = s). \end{aligned}$$

Correspondingly, we could estimate the fraction of spikes in cluster c that was expected to be assigned to other (i.e. wrong) clusters and obtained a generalized fraction of false negatives:

$$\begin{aligned} f^n_k &= \sum_{i\neq k} f^n(k;i) \\ &= \frac{1}{N_k}\sum_{i\neq k}\sum_{s\in i} P(C = k|S = s) \end{aligned}$$

The values of $P(C = c, S = s)$ were given by Gaussian distributions with parameters found by the expectation-maximization algorithm.

Confocal imaging and image analyisis

After the recording, the retina was immediately fixed with 4% paraformaldehyde (in 0.1 M PBS, 200 mM Sucrose) on the MEA chip for 1 hour. The retina was then rinsed several times with

0.1 M PBS and embedded with Vectashield (Vector Laboratories, UK) and sealed with a coverslip (Menzel Glaeser, Germany). Imaging was performed with a Leica SP5 confocal upright microscope supplied with a 25x / 0.95NA WD 2.5mm water immersion objective for an optimal signal collection focusing on 8x8 electrode arrays in 300x300 μm field of view. In each field, images (2048x2048 pixels) were acquired in z-stacks in tissue thickness 60-100 μm with optical slicing that corresponded to 30-50 image planes in each tissue volume. Acquisition parameter optimization revealed that a lateral resolution of 200 nm per pixel, just above the diffraction limit, and optical slicing of 550 nm provided an adequate trade-off between the level of image detail for morphological analysis and the acquisition time minimizing the risk of photo damage for long exposure. Microscope parameter optimization was performed using tools to increase the signal-to-noise ratio (SNR), including high number of frame averaging with an upper limit determined by safe levels of laser power to protect the tissue, and post image processing methods using deconvolution. In order to increase the image quality in varying depth locations in the highly scattering retina, which occur due to optical inhomogeneities, deconvolution using Richardson-Lucy algorithm was applied with several iterations; the number of iterations was chosen depending on the noise level and image blur, usually in the range from 3 to 10. In addition to the fluorescence signals in specific fields, large-field images including images of the electrode array were also acquired in order to enable the colocalization of images with RGC spiking activity.

In one Thy1 YFP/ChR2 retina, RGC somata were manually annotated in selected subfields where activity was recorded, and the confocal images of the RGC layer were spatially aligned with the estimated locations of detected events. To this end, the active area of one electrode was determined, and the remaining electrode locations were computed generating a regular grid using the 42 μm electrode spacing. The images and soma locations were then transformed into array coordinates, and spike locations were overlayed with the retinal image.

Results

The full analysis from raw data to sorted spikes, and subsequent quality control, involved several steps, which are described in the following sections. To present the relevant features of the data, and to evaluate the performance of the methods, we employed recordings from representative mouse retinas. Figure 1A illustrates a typical recording setup, with a flattened retina placed on the array.

Spatial event localization

In a first step, spikes were detected using a previously described threshold-based method that exploited dense sampling to improve detection performance, and was capable of estimating current source locations for detected spikes (Muthmann et al., 2015). This procedure yielded spatio-temporal event maps, where each event was identified by a time stamp, two spatial coordinates and its interpolated waveform (Figure 1B).

These maps revealed structures in the activity at a spatial resolution higher than that provided by the electrode arrangement of the MEA (Figure 1B,C). As expected for signals originating primarily from localized clusters of sodium channels at the axon initial segment, spikes were found in dense clusters surrounded by areas of low event density. The relationship between recorded signals and spike locations is illustrated in Figure 1C. The estimated locations of several spikes were marked by circles and the corresponding segments highlighted in the raw data traces from nearby electrodes. These examples show how spike locations are related to the spatial decay of the voltage peaks, for which we found a roughly inverse spatial decay (Figure 1D, cf. Pettersen and Einevoll 2008; Lindén et al. 2011; Mechler et al. 2011). Thus, on dense MEAs, location estimates of current sources could potentially aid spike sorting through spatial clustering. Yet, an inspection of spike waveforms even in a small area suggested the presence of multiple units (Figure 1E), indicating that spatial clustering alone was insufficient for reliable isolation of single unit activity, as also

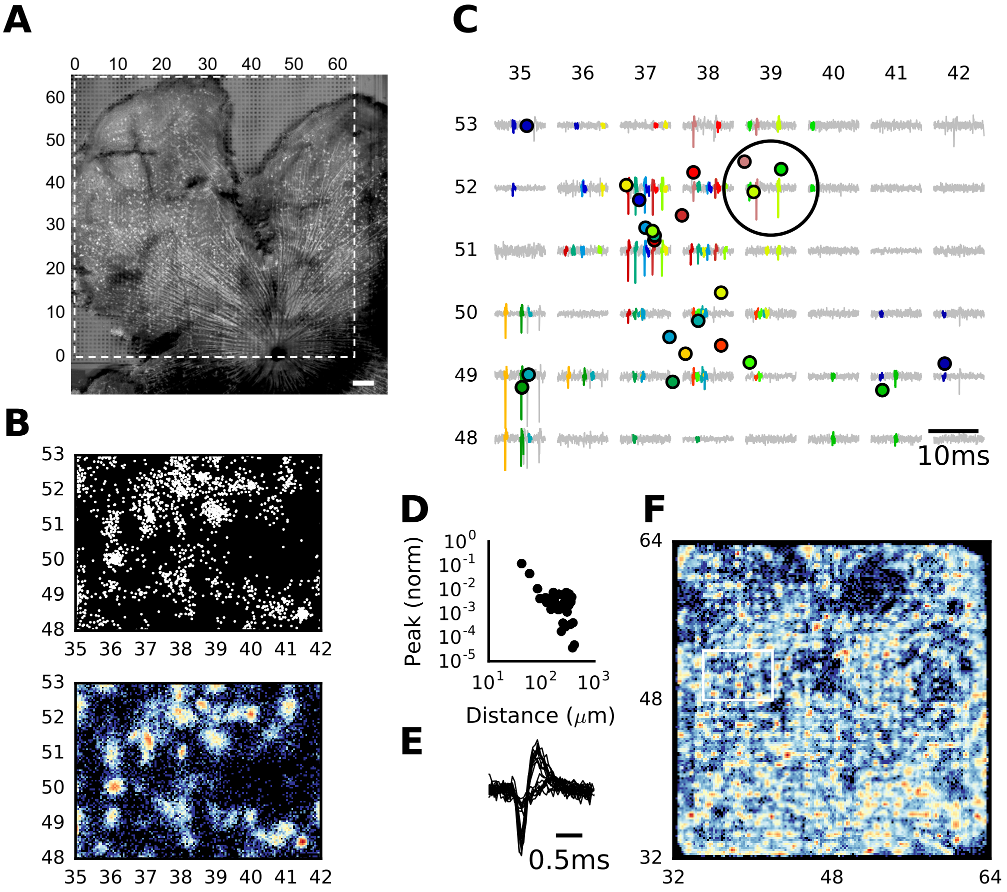

Figure 1 Spatial event localization reveals isolated spike clusters. **A**, Confocal image of a retina expressing YFP under the Thy1 promoter, placed on the array for recording. Electrodes can be seen as small squares in areas not covered by the retina. The active area of the array is indicated by dashed lines. Scale bar is 200 μm. **B**, All spikes detected during 10 s recording time plotted at their estimated locations (top), and spike counts detected during 2.5 minutes shown as density plot (bottom). Both plots reveal distinct clusters in space presumably originating from different neurons. **C**, Examples of several detected events, shown at their estimated locations (colored circles), and the corresponding episodes in the raw data (colored traces). Note only events localized within the visible area are highlighted. **D**, Average peak signal decay for detected events as a function of distance. The plot is suggestive of an inverse decay, as expected for electric potentials. **E**, Twenty randomly selected spike shapes for events localized within the area marked by the large circle in panel C, indicating the presence of signals from at least two different neurons at this location. **F**, Activity map for the whole recording. All data in panels B-F are from the same recording acquired with 24 kHz on 32x32 channels, panel A shows a different preparation.

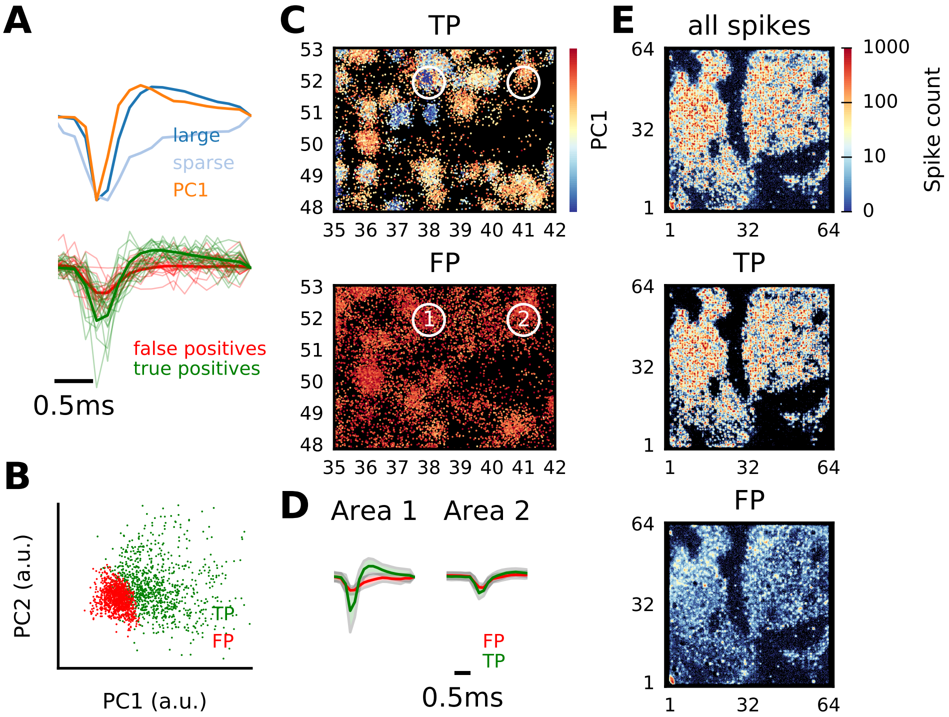

Figure 2 Classification of spike shapes. **A**, Average waveforms of events sampled from areas with low event density (“sparse”) and with high amplitude (“large”) samples, which were used to train the classifier (top). Example waveforms of events classified as true and false positives are shown below. **B**, Projections along the first two principal components (PCs) of waveforms classified as true and false positives (TP, green, and FP, red, respectively). **C**, Events classified as true (top) and false positives (bottom), at their estimated locations. Color indicates the projection along the first PC. **D**, Average waveforms of all TP and FP in the two circled areas in panel C. **E**, Spatial event density maps for a complete recording. Shown are all spikes (top), true (middle) and false positives (bottom). Data in this figure are from the same retina as Figure 1, but was recorded at 7 kHz.

shown in earlier work (Prentice et al., 2011). Moreover, since the detection was performed at a very low threshold to minimize false negatives, the activity maps still contained noise. For instance, events were detected in areas without tissue on the MEA (Figure 1F), so the ensuing analysis had to be resistant to such noise and should be able to cluster spikes based on a combination of location and waveform features.

Shape-based event filtering

The remaining noise in the activity maps could be reduced by increasing the detection threshold, or by applying additional shape criteria before acceptance. Yet since such heuristic strategies introduced arbitrary decision boundaries, we found that this generally increased the fraction of false negatives, for instance by removing events that showed clear light responses although they did not fulfill prespecified shape criteria. Therefore, a post hoc event selection was performed, which also helped in compensating for variations between preparations and improving recording quality.

To distinguish between true and false positives, examples of events with either high amplitudes or from areas with very low spike density were randomly chosen from the data and used to train a radial basis function Support Vector Machine (SVM) classifier. High amplitude events showed the typical biphasic spike waveform, which resembled the first principle component (PC) estimated from all events, while low density events lacked the repolarization (Figure 2A). This confirmed the premise that regions with very low spike density contained almost exclusively noise. A classifier trained on these examples separated events roughly, but not exactly along projections along the first PC (Figure 2B).

The comparison in Figure 2C showed that events classified as true positives were typically part of localized spatial density peaks, while false positives were more homogeneously distributed with low density (Figure 2C). True positives show clear, biphasic waveforms (see area 1 in Figure 2C,D). Yet the separation would still remain ambiguous for small events with amplitudes closer to the noise level, where events classified as false positives still showed spatial clustering (area 2 in Figure 2C,D), suggesting the presence of poorly detected current sources as well as other events related to neural activity such as strong synaptic currents (Muthmann et al., 2015). As shown in Figure 2E for a whole 64x64 channel recording, most of the spatial structure was retained in the map of spikes classified as true positives, while events in areas where no spikes were expected (e.g. optic disk, incisions) were correctly removed. On the other hand, the map of false positives showed weak spatial clustering in areas with high activity. This indicated that the spike record of some neurons with weak signals was most likely incomplete, and a further selection of events had to be performed after spike sorting (see below).

Combined spatial and shape-based clustering

The final step of the analysis was clustering the spikes into single units. Spike sorting traditionally relies on differences in waveforms caused by the locations of neurons relative to the recording electrodes. Our recordings provided spatio-temporal footprints for all events originating from fluctuations in electrical potentials recorded at several electrodes. These footprints could provide sufficient information to enable reliable source discrimination (Prentice et al., 2011). Yet, in practice this would require solving a complex assignment problem, which was computationally extremely challenging for data from thousands of channels (Rossant et al., 2016).

To solve this problem, we relied on a highly efficient event representation where a fast clustering algorithm could be deployed. Specifically, each event was described by its estimated location on the array, and by features extracted via PCA from the single interpolated waveform. Since spikes tended to be grouped into spatially dense clusters, the Mean Shift algorithm (Comaniciu and Meer, 2002) was particularly effective for clustering these activity maps. For spatially overlapping units, the additional waveform features then provided additional constraints for separation. In practice, we found that including the dominant two PCs was sufficient to successfully isolate single units, hence the full clustering and assignment task was reduced into a four dimensional clustering problem, which could be performed in just minutes for millions of events on a fast computer.

Figure 3 shows a complete sorted data set containing about 7.6 million events, which were separated into more than 2,500 clusters. Cluster sizes ranged from hundreds of spikes to several tens of thousands, corresponding to firing rates from 0.1 to 30Hz (areas of the circles in Figure 3A indicate the firing rate of each unit). A magnified view of a subset of clusters illustrates several important features of the data (Figure 3B). Spikes with clearly bi-phasic and large amplitude waveforms formed the most coherent spatial clusters, while smaller events form clusters that are spatially more spread out (compare average waveforms for the clusters). As also shown above, the strength of the first PC projection (PC1) for the events gave a good indication of their biphasic character (Figure 3C). The magnitudes of these projections were color-coded in Figure 3A for each cluster, and in panel B for each spike (note that the sign would be flipped such that larger values always coincided with more biphasic waveforms). Visual inspection showed that well detected units could be more reliably clustered, while cluster boundaries for spikes with weak signals were occasionally more ambiguous. Absence of a clear relationship between firing rate and PC1 suggested that while variable, the detection quality of individual neurons was not biased by the functional classification of the neuron, a result that will be substantiated below using light responses.

As already seen above, this suggested it was possible to discard a subset of poorly detected neurons with unobserved spikes below the detection threshold, for which spikes were eventually not reliably clustered. While a number of heuristic criteria could be used possible, an easy method was to

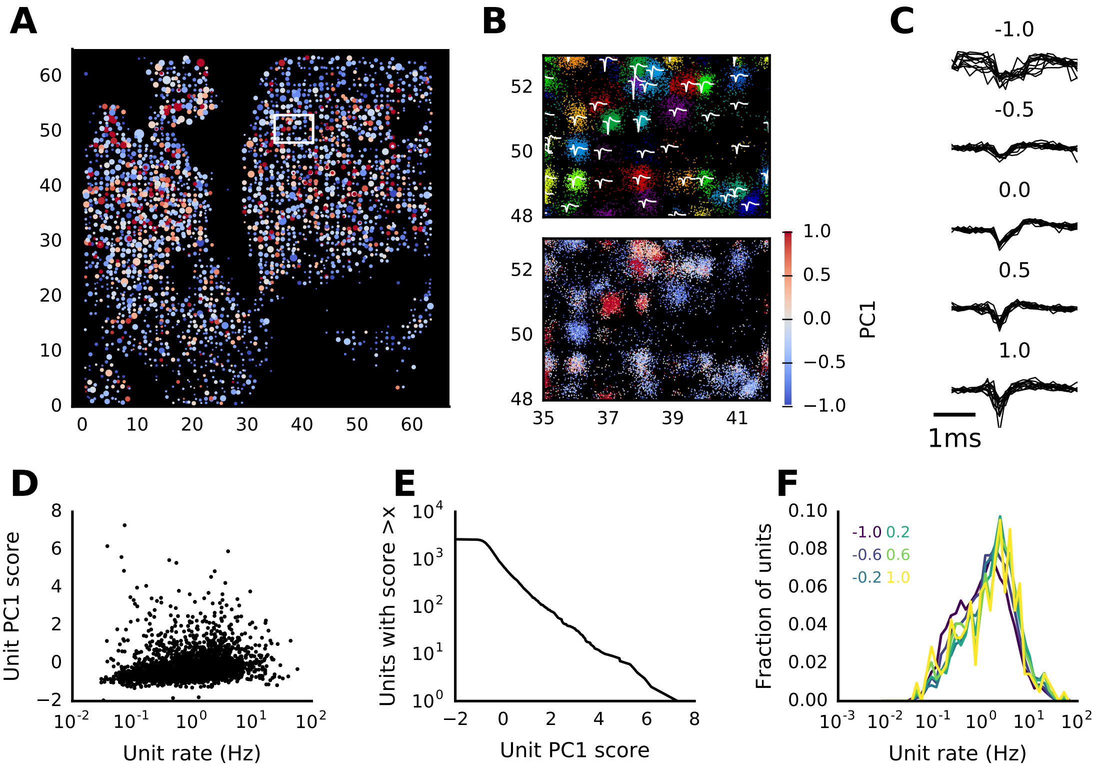

Figure 3 Clustering of all events into single units. **A**, Overview of all single units obtained by clustering a recording with 7.8 million detected events. A total of 2,623 clusters were formed, which are shown as circles at their estimated locations in array coordinates. Color indicates the coefficient (score) of the first principle component (PC), where larger values indicate a more biphasic waveform. **B**, Magnified view of a group of units (area in white rectangle in panel A). The top plot shows a subset spikes at their estimated locations (dots), and the average waveform associated with each unit. Each unit is shown in a different, unique color. Below the spikes are shown colored by their first PC scores. **C**, Examples of events with different first PC scores. **D**, There was no relationship between first PC score and firing rate for the units in this data set. **E**, Cumulative histogram of first PC scores. **F**, Firing rate histograms for units where progressively all units with an average first PC score below a set value were removed (thresholds given in legend). The overall shape of the histogram did not change when units with low scores were removed.

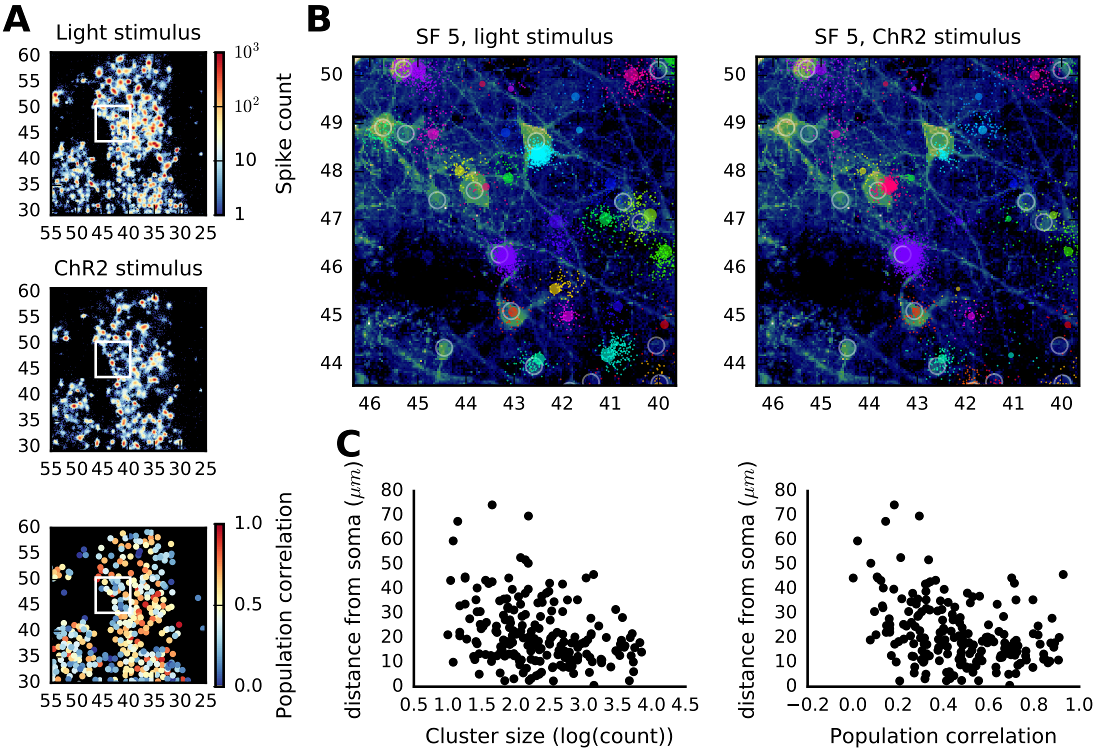

Figure 4 Comparison of optogenetically evoked spikes with anatomical imaging. **A**, Activity maps obtained during light stimulation (top) and ChR2 stimulation and complete block of neurotransmission (middle). The bottom graph shows the correlation of the activity of each unit with the overall population activity, which quantifies the responsiveness to optogenetic stimulation. **B**, Alignment of the activity with a confocal image. Individual spikes are shown as small dots, colored according to unit membership. Detected somata are highlighted by circles. The centroids the units are also shown as colored circles, with areas proportional to the spike counts. Cells expressed YFP under the Thy1 promoter, hence labeled neurons corresponds to those expressing ChR2. Most, but not all labeled neurons showed activity during normal light stimulation, and in some places activity could be seen in areas without a labeled neuron nearby (left). During optogenetic stimulation, strong activity was almost entirely restricted to areas with nearby labeled somata (right). **C**, Spatial distance to its closest soma for each unit, plotted against spike count during optogenetic stimulation (left). All strongly activated units had a corresponding soma within one electrode radius (42 μm). The same was the case for the population correlation where all units with strong correlation had a soma in close vicinity (right). Data in these graphs summarize an imaged area of 0.78 mm^2.

threshold the clusters by their average PC1 score. The distribution of these scores was right-skewed (Figure 3E), and a relatively conservative threshold of 0 still yielded about 1000 units for the illustrated data set. Importantly, excluding units did not change the shape of the firing rate distribution (Figure 3F), and was therefore unlikely to introduce a bias in the remaining cell sample.

Validation with anatomical images

Before we moved on to assess the quality of the spike sorted data, we tested whether we could indeed match the detected clusters with actual neurons. To this end, we used a transgenic mouse line expressing YFP and ChR2 under the Thy1 promoter in about half of all RGC (Raymond et al., 2008). The rationale was that this would enable us to stimulate spiking exclusively in the subset of visually identifiable RGCs because these cells also expressed ChR2. This would allow us to clearly establish correlates between single spike-sorted units and individual, visually identifiable RGCs.

We first compared the photoreceptor-driven activity recorded during normal light stimulation (ND 4.5, full field 0.5 Hz) with recordings in the presence of DNQX and L-AP4 and at a maximum irradiance of ND 2.2) to selectively evoke ChR2 mediated spikes (Figure 4A). The activity maps clearly showed that only a subset of all RGCs responded to optogenetic stimulation (Figure 4A, compare top and middle plots). For this particular dataset, we found 375 units with a firing rate of at least 0.5 Hz during light stimulation, but only 254 units during ChR2 stimulation. In addition, 77 units were significantly less active during light stimulation than during ChR2 stimulation, these were presumably neurons unresponsive to our light stimulus but nevertheless expressed ChR2.

Next we colocalized the activity with confocal micrographs showing the YFP labeled neurons (Figure 4B). In total, we analyzed an area of 0.78 mm^2, where 195 somata were manually annotated, and 211 units were detected (note that the sorting was performed on the combined recordings). An example of the alignment of activity and anatomical image is shown in Figure 4B, both for activity obtained during light stimulation (left) and ChR2 activation (right). While some units were clearly active in both conditions, others had more spikes in one than the other. Importantly, all units with significant activity during ChR2 stimulation were closely co-localized with a soma. Low levels of activity in neurons not expressing ChR2 were most likely due to intrinsically generated spontaneous activity, which was difficult to block.

To assess these results more quantitatively, we measured for each unit how well it followed the ChR2 activation, which was triggered by a sequence of random light flashes. As a measure, we determined the correlation of an individual unit's activity with the overall population activity. It was close to zero for spontaneously active units, but about 40% of all detected units had a correlation larger than 40%, indicating that they were well activated. Almost all of these units had a soma within less than one electrode radius ($<$42 μm, Figure 4C), and similarly, most units with high firing rates were localized close to one of the labeled somata. This showed that the signals detected and assigned to single units by our method indeed originated from single neurons, and that event localization was sufficiently precise to co-localize well detected units with their anatomical counterpart. The image alignments however also showed that some units had more spatially dispersed spikes, which was likely due to unfavorable signal to noise conditions.

Independent validation of sorted units and quality control

The list of units provided by the clustering step contained assignments for all putative spike events detected in the raw data, regardless of their signal amplitude and quality. Since detection was performed with a low threshold to minimize false negatives, in a final step automated post hoc quality control was performed to select the subset of well detected and clustered units for further analysis. Refractory violations are a frequently used tool, but these were virtually non-existent in our data and therefore not informative. Instead, we followed an approach proposed by Hill et al. (2011), which involved fitting an independent statistical model to the sorted data and estimating statistical sensitivity and specificity from the resulting posterior probabilities.

Due to the size of a whole data set, our strategy was to investigate each unit in turn, taking nearby units into account which could interfere with the unit in question. Specifically, a Gaussian mixture model with the same number of components as there were units, was used. It was fit to the spatial locations of all spikes and the waveform projections along the primary principal components (here two components were used, a larger number yielded almost identical results). Next, the fraction of false positives and negatives were computed directly from the posterior probabilities computed from the mixture model. A technicality arised when spike counts in the clusters were very unbalanced, where the fitting failed to detect the individual clusters. In such cases, the model was therefore fit to an equalized subset of the data.

Two typical outcomes of this procedure are illustrated in Figure 5, panels A and B, for relatively crowded areas on the array. The first was a unit with a very distinct waveform (blue), and three further units in close proximity, which were spatially well separated (Figure 5A). In this case, the blue unit was very well isolated when shape information alone (PCA), or combined shape and spatial information (X/Y+PCA) were used for clustering. The other units, however, had

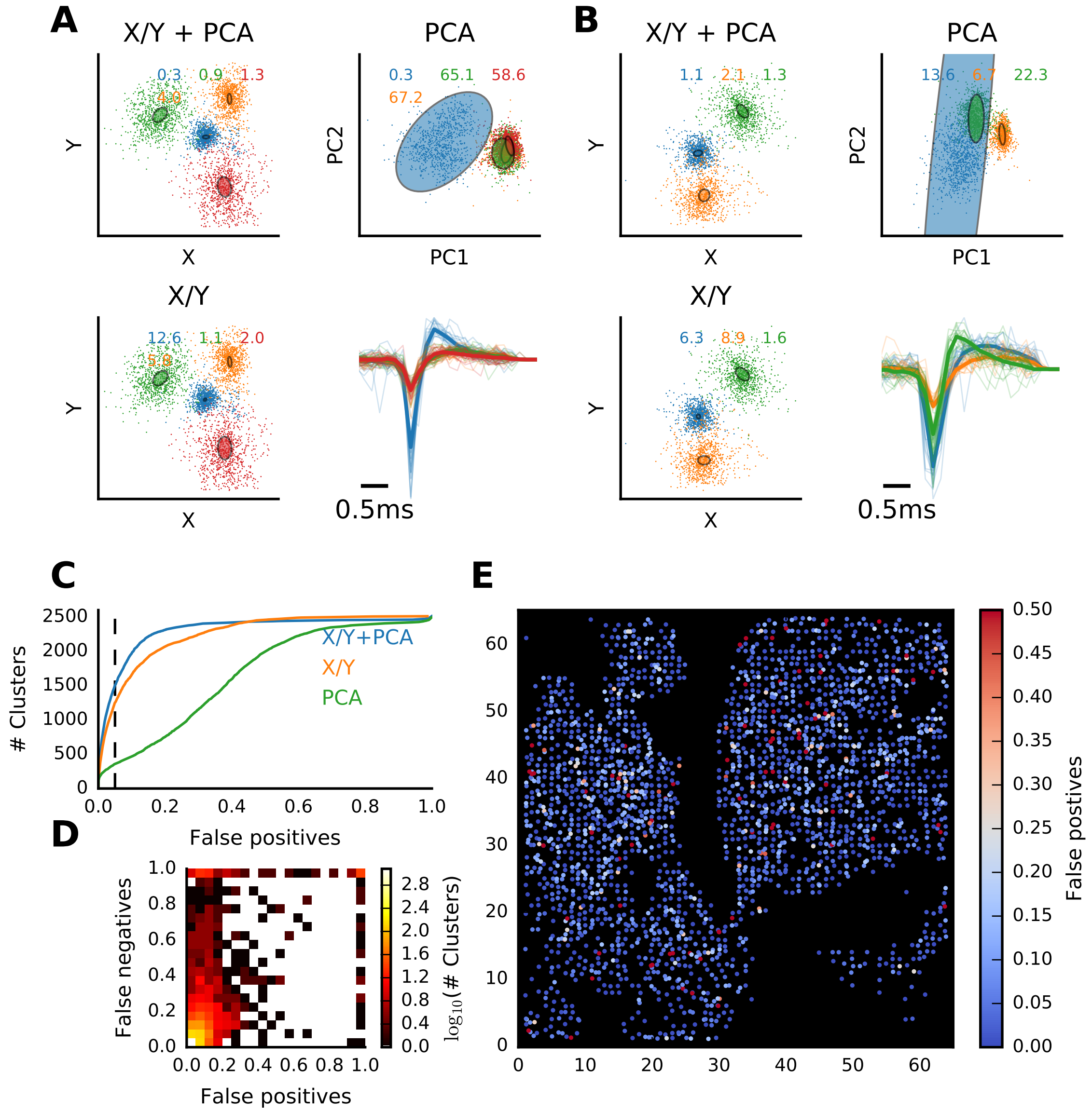

Figure 5 Quantitative assessment of sorting quality with Gaussian mixture models (GMM). **A**, **B**, Two examples of GMMs fit to groups of neighboring units. The fits were performed either using combined spatial and shape information (X/Y+PCA), shape information alone (PCA) or only spatial locations (X/Y). Spikes are colored to indicate the original cluster assignments. Numbers in each panels are percentages of false positives, as estimated from the posterior probabilities given by the GMM fit for each case. Small numbers indicate a good match of the original assignments with those predicted by the GMM. Each panel also also shows example spike waveforms and the unit average (thick line), using the same color scheme. **C**, Cumulative histogram of the fraction of false positives for all units in one recording. The vertical line marks 5%. **D**, Relationship between fractions of false positives and negatives for the whole data set. **E**, Spatial distribution of false positives for all units.

very similar spike shapes, but were well isolated when their locations are considered. The second example shows three units that were spatially quite well separated, but where the shapes were much more similar, so that it became difficult to perform sorting based solely on shape features (Figure 5B). Here, a combination of the two features yielded the best result for all units.

The results of this analysis performed on all units in a data set with 7.6 million spikes are summarized in Figure 5C-E. Each one of the approximately 2,500 units with a sufficiently high spike count (>150 spikes) took, in turn, the role of the blue unit in the examples above. For each unit, at least its closest neighbor, or all units within a radius of 31.5 μm (3/4 electrode pitch) were used in the fit.

When the combined spatial and shape information was used for quality control, 46% of the units (1210) had false positive and negative rates lower than 5%, and 22% (565 units) of these were below 1% (Figure 5C, blue line). These fractions decreased somewhat when the fit was only performed on spatial locations, and became much worse when only shape information was used (Figure 5C, red and green lines). In general, the estimated rates of false positives and negatives were correlated, but around 25% of all units had more than one neighboring unit, increasing the chance of mis-assignments and hence the false negative rates (Figure 5D). A spatial overview of these results showed that, as expected, poorly sorted units are primarily found in areas with dense activity (Figure 5E), where some units had a weak signal that prevented unambiguous classification.

Functional assessment of single unit activity

The recordings used in this paper contained responses to a series of full field flashes, which enabled us to directly evaluate light responses of the sorted units. Spike waveforms, inter-spike interval histograms, raster plots and peri-stimulus time histograms of units from units in a small patch of retina are shown in Figure 6. The example was chosen from a relatively crowded area of the recording (same data and area as in Figure 1) to illustrate potential problems caused by spatially overlapping units, and how units can be selected based on the quality measures introduced above.

The example shows units that were well separated in space, as well as several spatially overlapping units (Figure 6A). The percentages of false positives and negatives, given above the spike waveforms, indicated that in particular those units with large spikes (units 1, 3 and 4) were well separated. These units also appeared as relatively well defined clusters in the PCA projections (Figure 6C), or were spatially isolated as it was the case for unit 4. In contrast, units with small spikes, in particular those in regions with high activity, were less well clustered (units 5-7). This poor clustering was mainly due to events close to the cluster boundaries in PCA space. Unit 8 had small spikes, but was spatially relatively well isolated and had a distinct waveform, hence was well clustered too. Yet its spikes were quite close to detection threshold, and one would therefore expect that the spike record for this unit was incomplete.

Importantly, all units had rather distinct light responses, and there was no clear evidence of mixing of functional properties even for poorly sorted units. It is also notable that none of the units except perhaps unit 8 contained refractory period violations (Figure 6B). For further analysis however, a conservative choice would be to only retain the units with the best separation, and to discard the others. The visualization software we developed allows quick inspection of borderline cases, and flag units for exclusion.

The recording shown here was acquired at 24 kHz, and this high sampling rate indeed significantly improved clustering. We recorded the same area at 7 kHz, and reproduced the same clustering with almost identical light responses (data not illustrated). However, due to the reduced signal quality both spike localization and waveform features have more noise. As a result, only units 1, 3 and 4 were well isolated, hence the yield of well sorted units was lower in this case.

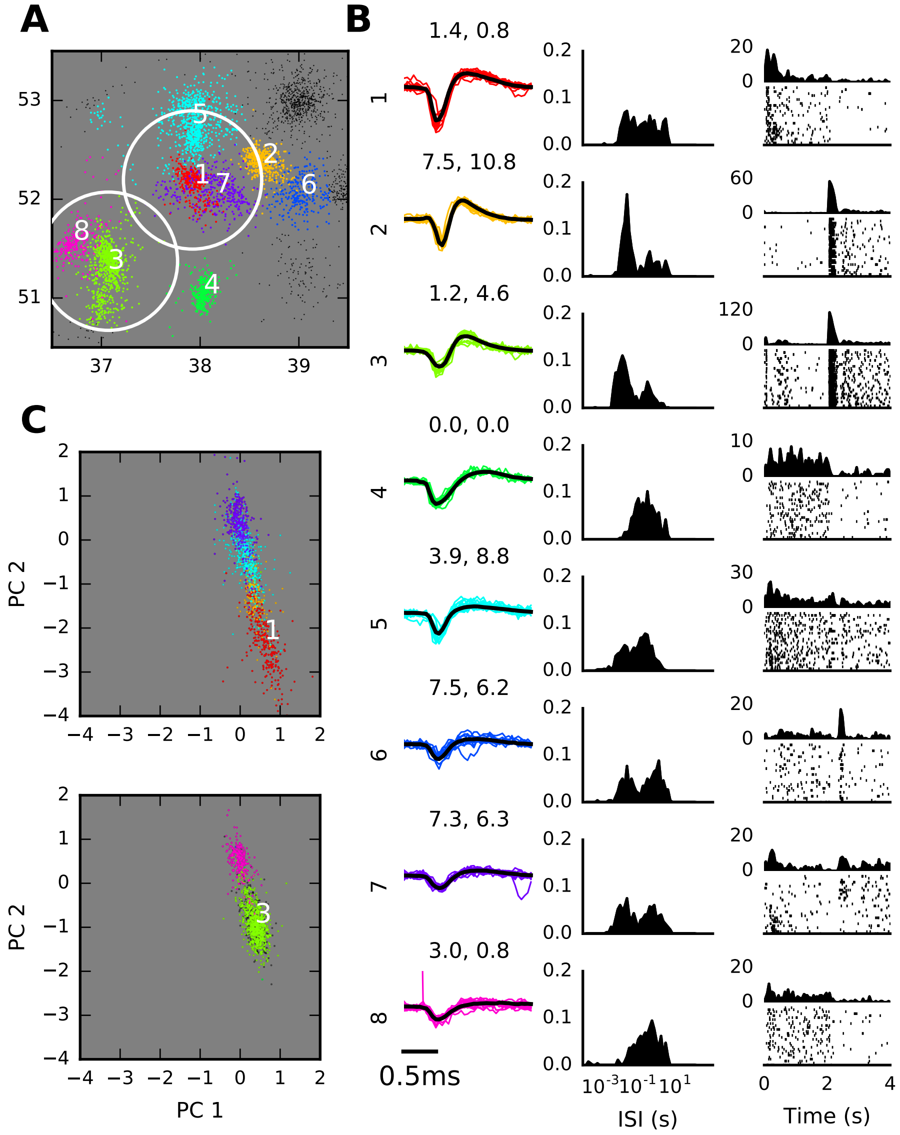

Figure 6 Functional characterization of spike sorted retinal ganglion cells. **A**, Event locations on a portion of the chip, as determined by interpolation between electrodes, and after clustering. Eight clusters with the largest spike amplitudes in this area are highlighted in different colors, while all remaining spikes are shown as black dots. Coordinates are in units of electrode distance (42 µm). **B**, Overview of the clusters highlighted in part A. The graphs show, for each cluster, individual spike waveforms and the median (black traces), the inter-spike interval histogram, and the raster and PSTH for full field stimulation. Percentages of false positives and negatives are given above the spike waveforms. **C**, All spikes in the two circled areas in panel A, with identical color coding, shown in the space of waveform principal components (PCA space). The same data as in Figure 1 are shown, acquisition rate was 24 kHz.

Discussion

Spike sorting is a critical step in the analysis of extracellular electrophysiological recordings. The erroneous assignment of spikes can have rather severe implications for the analysis and interpretation of multi neuron activity, so that methods using joint models of spike waveforms and neural activity may be required to avoid spurious or biased correlation estimates (Ventura and Gerkin, 2012; Muthmann et al., 2015). Working with high density recordings during this study has shed some light on potential limitations of purely shape-based spike sorting. A major issue is that it can be hard or even impossible to decide how many units the signal from a single electrode contains. If an electrode is positioned close to a group of neurons, one or perhaps two units with very strong signals may have sufficiently distinct waveforms to be separable, but the spatial event localization performed here also revealed that weaker, and hence much less distinguishable, signals often originate from different locations. Our method can cope well with such situations because spatial location estimates are sufficiently precise to disambiguate such cases, and in addition, signal interpolation reduces noise in detected signals, perhaps at the cost of rejecting some of the very weak signals. A main factor affecting sorting performance is thus the noise and bias in spatial localization, which depend on the signal quality (Muthmann et al., 2015).

Spatial event source localization and waveform interpolation enabled us to perform clustering in few dimensions, which makes this method extremely efficient. The complexity of the mean shift algorithm scales quadratic with the number of spikes, and the highly optimized version used here has a much better performance in typical situations. We developed a parallelized implementation, which allows sorting of millions of spikes in minutes (10 million spikes are sorted in about 8 minutes on a 12 core 2.6GHz Xeon workstation). This enables an iterative improvement of the two parameters of the clustering algorithm, before a final manual inspection may be performed.

The complete workflow consists of first performing event detection, followed by spatial localization, clustering, and finally an optional manual inspection. The former two are described in detail in Muthmann et al. (2015), and currently constitute the main bottleneck of the analysis chain. Detection currently runs at about 10x real time, and scales linearly with recording duration. The complexity of the spike localization scales linearly with the number of detected events, and runs roughly in real time for typical data sets. Both are however in principle parallelizable, which should improve performance significantly, a first implementation is currently under development.

In terms of sorting quality, we found that a four-dimensional representation of spikes, which includes the two main principal components of the averaged waveforms, was sufficient to obtain very well separated units, even at low acquisition rates of 7 kHz. Adding more dimensions only gave minor improvements, because higher principal components are partially redundant with spatial locations, and also increasingly affected by noise. Yet reducing the full signal on the array into a single, interpolated waveform and two coordinates inevitably discards information that may otherwise improve the outcome at the clustering step.

A different strategy for high density recordings, developed by Marre et al. (2012), is to estimate spatio-temporal templates, which are then used to identify spikes from each neuron (see also Dragas et al., 2014). This shifts the computational burden from spatial interpolation and source localization in our method to the deconvolution of spikes from raw data. We found that adding additional shape criteria in the detection stage could lead to false negatives, suggesting that templates will only yield reliable results if the firing rate of the neurons is high enough so they can be estimated sufficiently precisely. A third method, recently developed by Rossant et al. (2016) for high density *in vivo* probes, achieves to reduce dimensionality by masking out irrelevant parts of the data based on geometric constraints before clustering. This avoids an early discarding of information, as it is done in our method by signal interpolation and by Marre et al. (2012) by creating templates. It is however currently unclear how this method would perform on two-dimensional arrays, and whether it can be easily scaled up to thousands of channels.

Acknowledgements

This work was supported by the RENVISION (EU Grant FP7-FET-NBIS). MS was supported by the EuroSpin Erasmus Mundus programme and the EPSRC Doctoral Training Centre in Neuroinformatics (EP/F500385/1 and BB/F529254/1), and JOM by the EuroSPIN Erasmus Mundus programme and by NCBS/TIFR.

References

Ballini M, Muller J, Livi P, Chen Y, Frey U, Stettler A, Shadmani A, Viswam V, Jones IL, Jackel D, Radivojevic M, Lewandowska MK, Gong W, Fiscella M, Bakkum DJ, Heer F, Hierlemann A (2014) A 1024-channel CMOS microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro. *IEEE Journal of Solid-State Circuits* 49:2705–2719.

Barrett JM, Degenaar P, Sernagor E (2015) Blockade of pathological retinal ganglion cell hyperactivity improves optogenetically evoked light responses in rd1 mice. *Frontiers in Cellular Neuroscience* 9:330.

Berdondini L, van der Wal PD, Guenat O, de Rooij NF, Koudelka-Hep M, Seitz P, Kaufmann R, Metzler P, Blanc N, Rohr S (2005) High-density electrode array for imaging in vitro electrophysiological activity. *Biosensors & bioelectronics* 21:167–74.

Comaniciu D, Meer P (2002) Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24:603–619.

Dragas J, Jackel D, Hierlemann A, Franke F (2014) Complexity Optimisation and High-Throughput Low-Latency Hardware Implementation of a Multi-Electrode Spike-Sorting Algorithm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 4320:1–1.

Eversmann B, Jenkner M, Hofmann F, Paulus C, Brederlow R, Holzapfl B, Fromherz P, Merz M, Brenner M, Schreiter M, Gabl R, Plehnert K, Steinhauser M, Eckstein G, Schmitt-landsiedel D, Thewes R (2003) A 128 128 CMOS Biosensor Array for Extracellular Recording of Neural Activity. *IEEE Journal of Solid-State Circuits* 38:2306–2317.

Frey U, Sedivy J, Heer F, Pedron R, Ballini M, Mueller J, Bakkum D, Hafizovic S, Faraci FD, Greve F, Kirstein KU, Hierlemann A (2010) Switch-matrix-based high-density microelectrode array in CMOS technology. *IEEE Journal of Solid-State Circuits* 45:467–482.

Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsáki G (2000) Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *Journal of Neurophysiology* 84:401–414.

Hill DN, Mehta SB, Kleinfeld D (2011) Quality metrics to accompany spike sorting of extracellular signals. *Journal of Neuroscience* 31:8699–705.

Hutzler M, Lambacher A, Eversmann B, Jenkner M, Thewes R, Fromherz P (2006) High-resolution multitransistor array recording of electrical field potentials in cultured brain slices. *Journal of Neurophysiology* 96:1638–45.

Lewicki MS (1998) A review of methods for spike sorting: the detection and classification of neural action potentials. *Network* 9:R53–R78.

Lindén H, Tetzlaff T, Potjans TC, Pettersen KH, Grün S, Diesmann M, Einevoll GT (2011) Modeling the spatial reach of the LFP. *Neuron* 72:859–872.

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

Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy TE, Berry MJ (2012) Mapping a complete neural population in the retina. *Journal of Neuroscience* 32:14859–73.

Mechler F, Victor JD, Ohiorhenuan I, Schmid AM, Hu Q (2011) Three-dimensional localization of neurons in cortical tetrode recordings. *Journal of Neurophysiology* 106:828–848.

Müller J, Ballini M, Livi P, Chen Y, Radivojevic M, Shadmani A, Viswam V, Jones IL, Fiscella M, Diggelmann R, Stettler A, Frey U, Bakkum DJ, Hierlemann A, Muller J, Ballini M, Livi P, Chen Y, Radivojevic M, Shadmani A, Viswam V, Jones IL, Fiscella M, Diggelmann R, Stettler A, Frey U, Bakkum DJ, Hierlemann A (2015) High-resolution CMOS MEA platform to study neurons at subcellular, cellular, and network levels. *Lab on a Chip* 15:2767–2780.

Muthmann JO, Amin H, Sernagor E, Maccione A, Panas D, Berdondini L, Bhalla US, Hennig MH (2015) Spike Detection for Large Neural Populations Using High Density Multielectrode Arrays. *Frontiers in Neuroinformatics* 9:1–21.

Obien MEJ, Deligkaris K, Bullmann T, Bakkum DJ, Frey U (2015) Revealing neuronal function through microelectrode array recordings. *Frontiers in Neuroscience* 9:423.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pettersen KH, Einevoll GT (2008) Amplitude variability and extracellular low-pass filtering of neuronal spikes. *Biophysical Journal* 94:784–802.

Prentice JS, Homann J, Simmons KD, Tkačik G, Balasubramanian V, Nelson PC (2011) Fast, scalable, Bayesian spike identification for multi-electrode arrays. *PloS One* 6:e19884.

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation* 16:1661–87.

Raymond ID, Vila A, Huynh UCN, Brecha NC (2008) Cyan fluorescent protein expression in ganglion and amacrine cells in a thy1-CFP transgenic mouse retina. *Molecular Vision* 14:1559–1574.

Rey HG, Pedreira C, Quian Quiroga R (2015) Past, present and future of spike sorting techniques. *Brain Research Bulletin* 119:106–117.

Rossant C, Kadir SN, Goodman DFM, Schulman J, Hunter MLD, Saleem AB, Grosmark A, Belluscio M, Denfield GH, Ecker AS, Tolias AS, Solomon S, Buzsáki G, Carandini M, Harris KD (2016) Spike sorting for large, dense electrode arrays. *Nature Neuroscience* 19:634–641.

Ventura V, Gerkin RC (2012) Accurately estimating neuronal correlation requires a new spike-sorting paradigm. *Proceedings of the National Academy of Sciences* 109:7230–7235.