nature COMMUNICATIONS

ARTICLE

Received 5 Dec 2014 | Accepted 16 May 2015 | Published 25 Jun 2015

DOI: 10.1038/ncomms8516 OPEN

Superstatistical analysis and modelling of heterogeneous random walks

Claus Metzner[1], Christoph Mark[1], Julian Steinwachs[1], Lena Lautscham[1], Franz Stadler[1] & Ben Fabry[1]

Stochastic time series are ubiquitous in nature. In particular, random walks with time-varying statistical properties are found in many scientific disciplines. Here we present a superstatistical approach to analyse and model such heterogeneous random walks. The time-dependent statistical parameters can be extracted from measured random walk trajectories with a Bayesian method of sequential inference. The distributions and correlations of these parameters reveal subtle features of the random process that are not captured by conventional measures, such as the mean-squared displacement or the step width distribution. We apply our new approach to migration trajectories of tumour cells in two and three dimensions, and demonstrate the superior ability of the superstatistical method to discriminate cell migration strategies in different environments. Finally, we show how the resulting insights can be used to design simple and meaningful models of the underlying random processes.

[1] Department of Physics, Biophysics Group, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91052, Germany. Correspondence and requests for materials should be addressed to C.M. (email: claus.metzner@gmail.com).

Stochastic time series, here used synonymously with random walks, play an important role in earth- and life sciences, technology, medicine and economics. Most of these disciplines deal with complex systems in which multiple hierarchical processes are interacting at different timescales. Systems with this level of complexity are likely to change their statistical properties as a function of time, resulting in heterogeneous time series. It is therefore surprising that only few tools are available for the analysis and characterization of such time-varying random walks. Some of these tools are used in finance[1–3], mainly with the goal of forecasting. In science, heterogeneous time series have been successfully described by Hidden Markov models[4]. However, systems with continuously time-varying statistics cannot be adequately modelled by a few discrete hidden states.

Owing to this lack of appropriate tools, many studies are still relying on conventional evaluation methods that were designed for simple physical systems. The most frequently used statistical measures for random walks, in particular the step width distribution (SWD), the mean-squared displacement (MSD) and the velocity autocorrelation function, are implicitly assuming that the stochastic process can be globally described by a few characteristic parameters, such as a constant variance and a constant correlation time.

We demonstrate in this paper that the application of these conventional methods to heterogeneous random walks generates 'anomalous' results, such as non-Gaussian SWDs or power-law MSDs with fractional exponents[5–7]. These anomalies emerge inevitably from the temporal averaging over changing local statistics during the evaluation period (Supplementary Note 1), and therefore do not provide meaningful insights into the random walk apart from its heterogeneous nature. Moreover, these temporally averaging measures may remain unchanged even if the experimental conditions are significantly altered. This lack of sensitivity points to a fundamental limitation of conventional statistical methods for analysing heterogeneous processes. SWD, MSD and autocorrelation function average over the successive statistical parameters of the heterogeneous random walk, instead of using the parameter dynamics as a rich additional source of information.

In this study, we propose a superstatistical framework for modelling and analysing heterogeneous random walks. The term superstatistics refers to the superposition of several different stochastic processes[8–11]. Accordingly, we describe the time series locally by a homogeneous random walk model with a minimum number of statistical parameters. In the case of cell migration, we use an autoregressive process of first order (AR-1) with a persistence parameter q and an activity parameter a. These parameters (q_t,a_t) are allowed to change with every time step of the random walk. By this way, heterogeneous time series of arbitrary complexity can be described (Supplementary Note 2).

We provide a new sequential Bayesian method to infer the time-dependent parameters from measured random walk trajectories. In contrast to conventional maximum likelihood parameter estimation within a sliding time window, our method can handle both gradual and abrupt changes of the parameters. As a Bayesian method, it provides not only point estimates but also their confidence intervals. After extraction of (q_t,a_t) from the measurements, the statistical properties of the time-dependent parameters can be subsequently analysed by computing the temporally averaged joint posterior distribution $p(q,a)$, the temporal auto-correlations $C_{qq}(\Delta t)$ and $C_{aa}(\Delta t)$, and the cross-correlations $C_{qa}(\Delta t)$.

In this paper, we use the migration of individual tumour cells as a case study of superstatistical analysis. Cell migration plays an essential role in many fundamental biological processes, such as embryogenesis, tissue repair or cancer development[12–14]. Anomalous features of cellular random walks have been reported by several groups, and a variety of models have been proposed in the literature to account for those anomalies[5,7,15–18].

We demonstrate that anomalies of conventional statistical measures to describe cell migration are attributable to fluctuations of migration persistence q and activity a. Moreover, the joint distribution of persistence and activity, $p(q,a)$, and the auto- and cross-correlations $C_{ij}(\Delta t)$ of these two parameters provide characteristic fingerprints of the underlying random walks. Unlike globally averaging statistical measures, a superstatistical analysis can clearly resolve the effects of different environments on cell migration, such as migration in a three-dimensional (3D) collagen network versus migration on a planar 2D culture dish. Furthermore, by observing individual cells in microfabricated 1D channel structures with varying diameter, we demonstrate that the temporal changes of the (q_t,a_t)-parameters are directly associated with different local microenvironments that the cells experience along their migration path. Finally, we show how the extracted statistical properties of the time-dependent parameters can be used to construct simplified models that reproduce all key features of the data, including the non-Gaussian SWD and power-law MSD. While other types of models have also successfully reproduced these anomalous features, for example, using fractional diffusion equations[7] or integro-differential equations with complex memory kernels[19], the superstatistical framework achieves this with the simplest persistent random walk model (the two-parameter AR-1 process), extended by the temporal variations of the two parameters (persistence and activity).

Results

Cell migration in 2D and 3D. We study the migration of the breast carcinoma cell line MDA-MB-231 in a 3D collagen gel and on a tissue culture-treated 2D plastic surface, either uncoated and or coated with the adhesion ligand fibronectin. Three-dimensional cell positions within the random fibre network of a collagen gel (Fig. 1a,b) are detected by analysing the characteristic refraction pattern (Fig. 1b inset) around the cell nucleus. From the individual cell trajectories (Fig. 1c), we compute momentary migration properties, such as cell speed versus time (Fig. 1c inset). Since the gel has a free upper surface and thus a lower effective stiffness in the z-direction, cells react with a more pronounced horizontal (x–y direction) alignment and motion, in agreement with theoretical predications based on active cellular mechanosensing mechanisms[20]. Therefore, only the x–y coordinates are used for comparing 2D and 3D migration.

Globally averaging statistical measures. For each individual cell trajectory, we compute the SWD, defined as the probability $p(\Delta x,\Delta t)$ that the cell changes its x-coordinate by Δx within a lag time interval Δt, as well as the MSD, defined as $r^2(\Delta t) = \langle(\mathbf{r}(t+\Delta t)-\mathbf{r}(t))^2\rangle_{t,e}$, where $\langle\rangle_{t,e}$ indicates temporal and subsequent ensemble averaging over the different individual cells of the same migration environment.

Regardless of environment, the SWD shows a leptocurtic, approximately exponential shape (Fig. 5a inset and Supplementary Note 3). For lag times below 500 min, the MSD can be approximated by power laws (Fig. 5a) with a fractional exponent of 1.3 in the cases of 3D collagen and uncoated 2D plastic, but with a larger exponent of 1.7 in the case of fibronectin-coated 2D plastic. It is remarkable that the SWD and MSD are practically indistinguishable for migration in 3D collagen and on uncoated 2D plastic, even though these environments require different migration strategies.

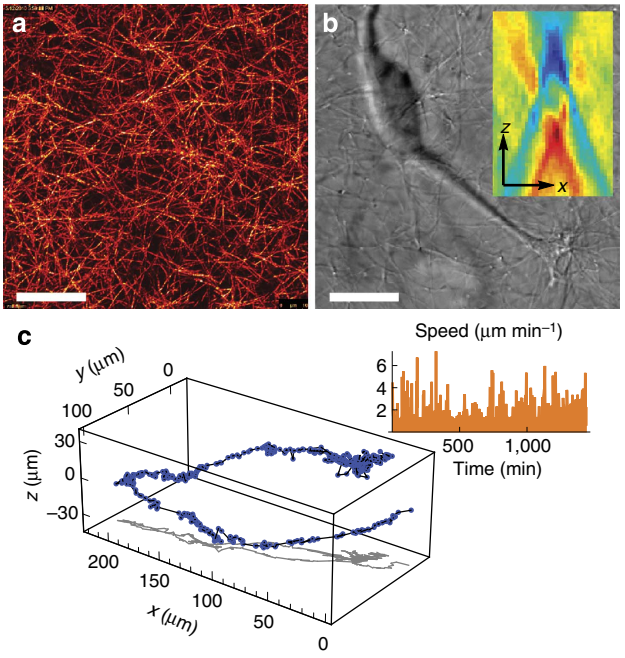


Figure 1 | Tracking and analysis of cells migrating in 3D collagen networks. (**a**) Confocal image of a collagen gel. (**b**) Bright-field image of an MDA-MB-231 breast carcinoma cell that has migrated into the bulk of the collagen to a depth of 200 μm. Scale bars, 20 μm. Inset: the characteristic light intensity profile (*z*–*x* plane) around the cell nucleus is used to track the cell position within the gel with an accuracy of 2 μm (r.m.s.). (**c**) Example of a 3D cell trajectory, sampled at 2.5 min time intervals. Inset: momentary speed as a function of time.

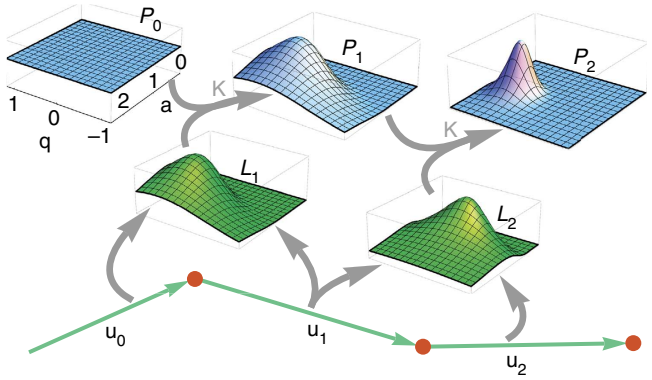


Figure 2 | Bayesian inference of time-dependent random walk parameters. From two successive displacement vectors $\mathbf{u}_0$ and $\mathbf{u}_1$, the likelihood $L_1(p,a)$ (green) of the parameters can be computed. This distribution is multiplied (grey) with the prior guess $P_0(p,a)$ (blue). *K* denotes a transformation that accounts for temporal parameter evolution. This process is iterated in forward and backward (not shown) time direction, and the priors are combined.

Within collagen, cells assume a pronounced elongated shape and typically form a path-finding long and thin protrusion that can extend over >100 μm (Supplementary Movies 1 and 2; ref. 21). The directionally persistent trajectory of the cells is mainly defined by the contour of this long protrusion, resembling the movement of a needle in an array of obstacles[22]. However, cells can also pull themselves along bundles of collagen fibres in a process known as contact guidance[23,24]. Occasionally, encounters with obstacles or small pores in the disordered collagen network can force the cell to withdraw or change directions (Supplementary Movie 2). On planar surfaces by contrast, the cells spread and assume a flat, irregular shape. They also polarize and move preferentially along their polarization axis (Supplementary Movie 3), but they cannot take advantage of external cues to keep a persistent migration direction.

Despite these diverging migration modes, the net spatial advancement of MDA-MB-231 cells over time is similar in both environments. Therefore, the SWD and MSD for migration in 3D collagen and on uncoated 2D plastic are nearly identical. On fibronectin-coated 2D plastic, the cells migrate more slowly but with a higher directional persistence (Supplementary Movie 4). Over time, this leads to a larger net spatial advancement compared with uncoated plastic. Accordingly, the MSD shows a higher fractional exponent of 1.7, and the SWD broadens (Fig. 5a).

Bayesian inference of time-dependent parameters. For the superstatistical analysis of the data, we first compute for each cell trajectory $\{\mathbf{r}_t=(x_t,y_t)\}$ the vectorial displacements $\mathbf{u}_t=\mathbf{r}_t-\mathbf{r}_{t-1}$ for each measurement time step $\delta t=5$ min. The statistical relationship between two successive displacements is described by a 2D first-order autoregressive process (AR-1) defined by

$$\mathbf{u}_t = q_t\mathbf{u}_{t-1} + a_t\mathbf{n}_t. \quad (1)$$

This process is equivalent to a persistent random walk or a time-discrete Ornstein–Uhlenbeck process. The parameter $q_t\in[-1,+1]$ describes the local persistence of the random walk, with $q_t=-1$ corresponding to anti-persistent motion, $q_t=0$ to non-persistent diffusive motion and $q_t=+1$ to persistent motion. The parameter $a_t\in[0,\infty]$ describes the local activity (noise amplitude) and sets the spatial scale of the random walk. Together, the two parameters determine the variance of the displacements according to $\mathrm{var}(u)=a^2/(1-q^2)$. The vector $\boldsymbol{n}_t=(n_{xt},n_{yt})$ is normally distributed, uncorrelated random noise with unit variance.

To extract the time-dependent joint probability density $P(q_t,a_t)$ of the parameters q_t and a_t from a sequence of displacements $\mathbf{u}_t$, we use sequential Bayesian updating. We start at time $t=0$ with a flat prior distribution $P_0(q,a)$ (see P_0 in Fig. 2), which can be interpreted as a 'first guess' about the parameter values. From the measured successive displacements $\mathbf{u}_0$ and $\mathbf{u}_1$, we compute the likelihood distribution $L_1(q,a)$ (see L_1 in Fig. 2), which provides a first information about probable parameter values.

The prior distribution P_0 and the likelihood distribution L_1 are multiplied to obtain the posterior distribution P_0L_1, which updates our guess of the parameter values for the next time step. In the case of a temporally homogeneous process with constant parameters, iterative multiplication of the posterior distributions with the likelihood distributions, $P_t=P_{t-1}L_t$ (Fig. 2), would yield an increasingly accurate estimate of the parameter values. For heterogeneous processes, however, the possibility of changing parameters has to be taken into account. This is achieved by a transformation K of the posterior distribution, $P_t=K(P_{t-1}L_t)$. The transformation K (blurring and preventing the posterior distribution to fall below a small cutoff value) is chosen such that both gradual and abrupt parameter changes can be identified (see Methods section). Finally, we perform the same sequential parameter inference in the reverse time direction (not shown in Fig. 2) and combine both distributions.

We validate this method by simulating random walk trajectories from prescribed stepwise (Fig. 3a) or gradually (Fig. 3c) changing parameter sequences $\{(q_t,a_t)\}$. We then reconstruct the parameter sequences from the simulated trajectories by sequential Bayesian inference. The mean values

of the posterior distributions fluctuate around the 'true' parameter values, but follow the prescribed time evolution closely, both for abrupt (Fig. 3a) and gradual (Fig. 3c) parameter changes. We also find that the Bayesian method is superior to a maximum likelihood estimation with a sliding time window. The maximum likelihood estimation method cannot handle abrupt and gradual parameter changes equally well, and the user must find a compromise between long time windows that wash out sudden parameter jumps and short windows that lead to noisy results (Supplementary Note 4).

Heterogeneity of measured random walks. We next apply the Bayesian inference method to measured cell trajectories. An example for the parameter evolution of a cell migrating on uncoated 2D plastic is shown in Fig. 4. We find large variations of cell behaviour, both with time (Fig. 4a,b) and between individual cells (Fig. 4c). By plotting the cell activity versus persistence for all time points, we further find that individual cells can occupy different regions of the (q,a) parameter plane (Fig. 4c). Some cells remain in a small compact region of the (q,a)-plane during the entire measurement period (brown), whereas others jump between disjunct subregions (green) or continuously change their parameters over time (Fig. 4c).

Superstatistical data evaluation. *Joint probability distributions*. We average the posterior distributions $p(q,a)$ for all time points and all cells measured in the same environment (Fig. 5b). In contrast to MSD and SWD, the ensemble-averaged posterior distributions show large differences between all three environments. The peak position of the distribution shows the lowest persistence and highest activity for collagen, and the highest persistence and lowest activity for fibronectin-coated plastic. Moreover, the spread of the distributions indicates that migration in collagen gels is more heterogeneous compared with migration on plastic. The $p(q,a)$ distributions thus provide characteristic 'fingerprints' of the migration environments that can be used for automatic trajectory classification. In a 'leave-one-out' cross-validation, we were able to assign ~90% of the cell trajectories to the correct environment (see Methods section).

Parameter correlations. The auto- and cross-correlations of the time-dependent parameters q_t and a_t reveal even larger differences between migration strategies in 2D versus 3D environments. Auto-correlation times are noticeably longer in a 3D environment (Fig. 5c), where the local biopolymer fibre configuration provides a guiding or trapping microstructure that influences a given migration mode for long time periods. Large differences between different environments are also visible in the cross-correlations of the time-dependent parameters (Fig. 5d). On

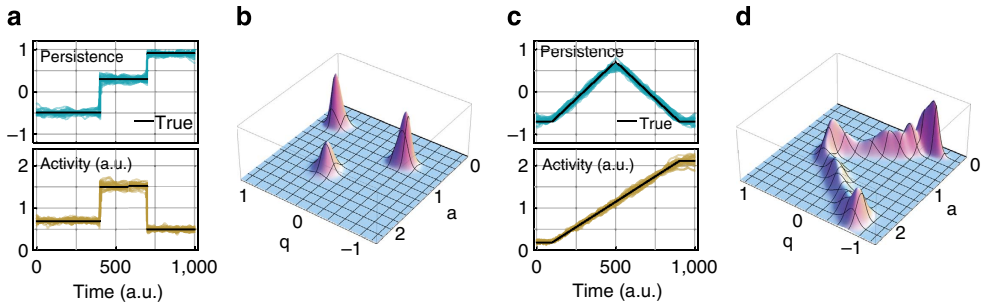


Figure 3 | Validation of Bayesian parameter inference with simulated data. The Bayesian method can reliably extract both abrupt (**a**,**b**) and gradual (**c**,**d**) parameter changes. (**a**,**c**) The prescribed parameter evolution (black) and reconstructions of persistence (blue), and activity (yellow) from multiple simulations. (**b**,**d**) The time-averaged joint posterior distributions of the parameters.

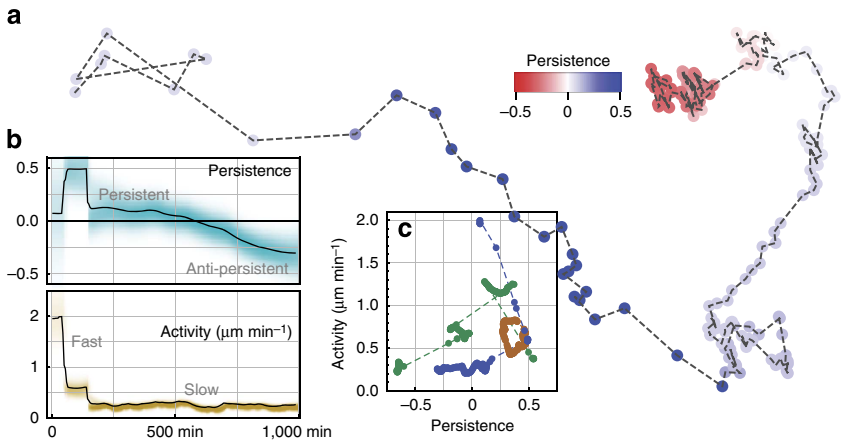


Figure 4 | Temporal heterogeneity of MDA-MB-231 tumour cell migration on uncoated plastic. (**a**) Example cell trajectory, with colours representing the posterior mean of the momentary persistence q_t according to the colour bar. (**b**) Persistence (top) and activity (bottom) of the same cell as a function of time. Shading intensity is proportional to the probability density distribution. The cell starts in a highly active and persistent state, switches within 200 min to a more inactive but still persistent state and then gradually changes from persistent (blue parts of trajectory) to anti-persistent (red parts) behaviour during the following 800 min. (**c**) Activity versus persistence ((q,a)-plane) for three individual cells (green, brown and blue) of the same type, migrating on uncoated plastic. The blue cell is the same cell shown in **a** and **b**. The dashed lines connect subsequent sampling points.

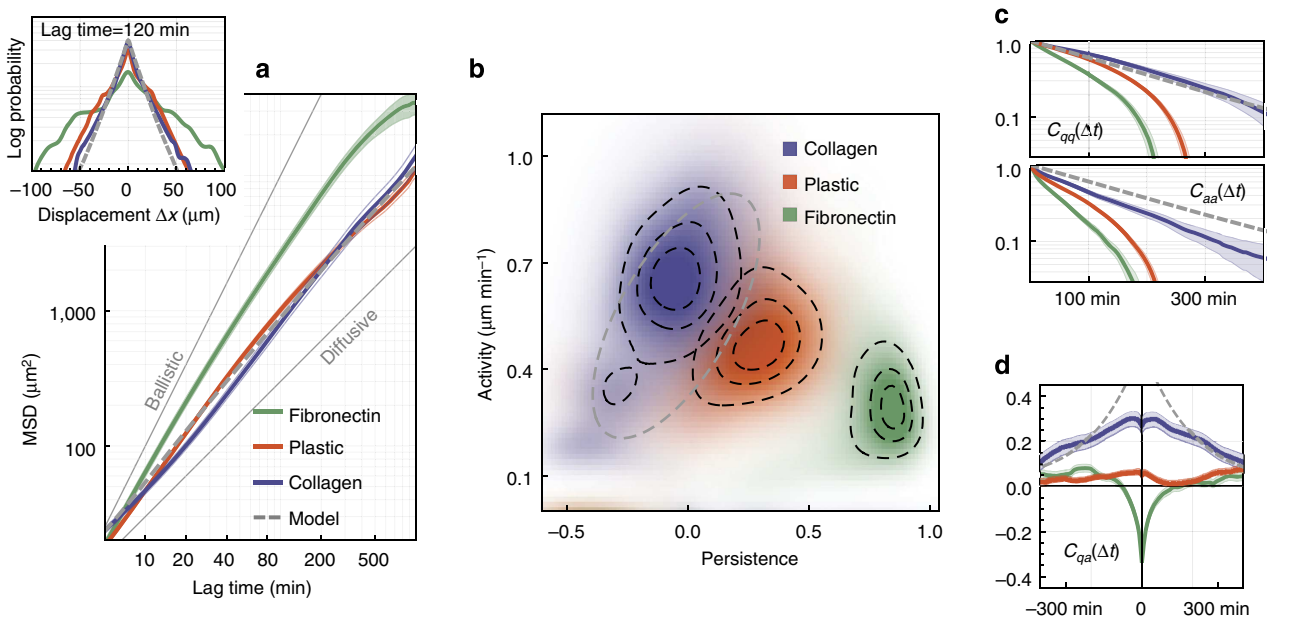


Figure 5 | Statistical evaluation of cell migration data. Conventional (**a**) and superstatistical (**b**–**d**) evaluation of migration data, ensemble-averaged over all cells in the same environment. The MSDs (**a**, main) grow superdiffusively with lag time, according to power laws with exponents 1.3 in collagen ($n=65$ cells from five experiments) and on uncoated plastic ($n=177$ cells from eight experiments), and 1.7 on fibronectin-coated plastic ($n=69$ cells from three experiments). The thin lines around each MSD curve indicate the s.e.m. (obtained with the bootstrap method). The SWDs are close to exponential for a lag time of 120 min (**a**, inset), as well as for other measured lag times (Supplementary Note 3). MSD and SWD show no differences between migration in collagen and on uncoated plastic. By contrast, large differences between all three environments are seen in the time-averaged joint parameter distributions (**b**), and also in the auto-correlations (**c**), and cross-correlations (**d**) of the parameters. Dashed black lines in **b** represent the 10, 25 and 50% credible regions. Shading in **c** and **d** corresponds to 1 s.e.m.. Dashed grey lines in **a**–**d** correspond to the superstatistical model of migration in collagen gels.

fibronectin-coated plastic, persistence and activity are negatively correlated for up to 100 min. This is consistent with the long-known observation that on highly adhesive surfaces, cells maintain persistent motion by performing sequences of small steps along the same direction[25]. The continuous gliding motion is not seen on less adhesive, uncoated plastic surfaces. Instead, we observe a weakly positive cross-correlations between q_t and a_t. In collagen, we find strong positive correlations between q_t and a_t, consistent with the observation that cells intermittently cover large distances with high directional persistence guided by long protrusions (Supplementary Movies 1 and 2).

Note that the activity parameter a_t should not be interpreted literally as the momentary cell speed u_t, but as a scale parameter that—together with q_t—determines the most probable value of the cell speed. To clarify this point, we also investigate the correlation between persistence q_t and momentary cell speed u_t. For migration on coated and uncoated plastic surfaces, we find a positive correlation between q_t and u_t (Supplementary Note 6). A similar relationship has been reported for a variety of different cell types migrating on fibronectin-coated surfaces[26]. In collagen, however, persistence and cell migration speed are uncorrelated (Supplementary Note 6).

Effect of local microenvironment. In the previous section, we have tacitly assumed that the local microenvironment has an immediate effect on migration persistence and activity. To test this assumption, we use a microstructured environment and measure cell migration through a linear (1D) array of sequentially narrowing channels and wider chambers. After extracting the time-dependent parameters q_t and a_t from individual cell trajectories (Fig. 6a), we plot q_x (Fig. 6b) and a_x (Fig. 6c) versus the x-position.

The precise migration mechanism of different cell types through such environments is not well understood and may involve integrin-mediated adhesion-dependent[27] or adhesion-independent[28] strategies. Regardless of the migration mechanism, our microstructured environment forces the cells to adapt to different degrees of confinement in rapid succession. A cell that enters a channel first has to polarize and deform its nucleus. It can then transit the channel with high persistence and activity. When the cell nucleus exits the narrow channel and enters the wider chamber, persistence and activity decrease markedly. Thus, the superstatistical migration parameters are strongly correlated with the local properties of the environment.

Superstatistical modelling. We construct a series of simple models of cell migration that approximate the statistical properties of q_t and a_t found in the data. All models are based on an AR-1 process. The superstatistical parameters q_t and a_t switch to new values, drawn from fixed distribution $p_{\mathrm{model}}(q,a)$, after exponentially distributed time intervals with mean value T_{model}. This regime-switching approach leads to exponentially decaying auto-correlations of the parameters with correlation time T_{model}. We choose $T_{\mathrm{model}}=200$ min taken from migration experiments in collagen (Fig. 5c). The parameter distribution $p_{\mathrm{model}}(q,a)$ is modelled as a bivariate Gaussian, centred at the main peak of the experimentally observed distribution $p(q,a)$ (Fig. 5b).

We first consider the limit of zero variance for $p(q,a)$, which corresponds to a homogeneous correlated random walk with constant q and a. In this case, the MSD is crossing over from a ballistic (slope 2) to a diffusive (slope 1) behaviour at a specific lag time that depends only on the persistence q. Increasing the variance of q generates a continuous mixture of crossover times, and the MSD starts to resemble a power law (Supplementary

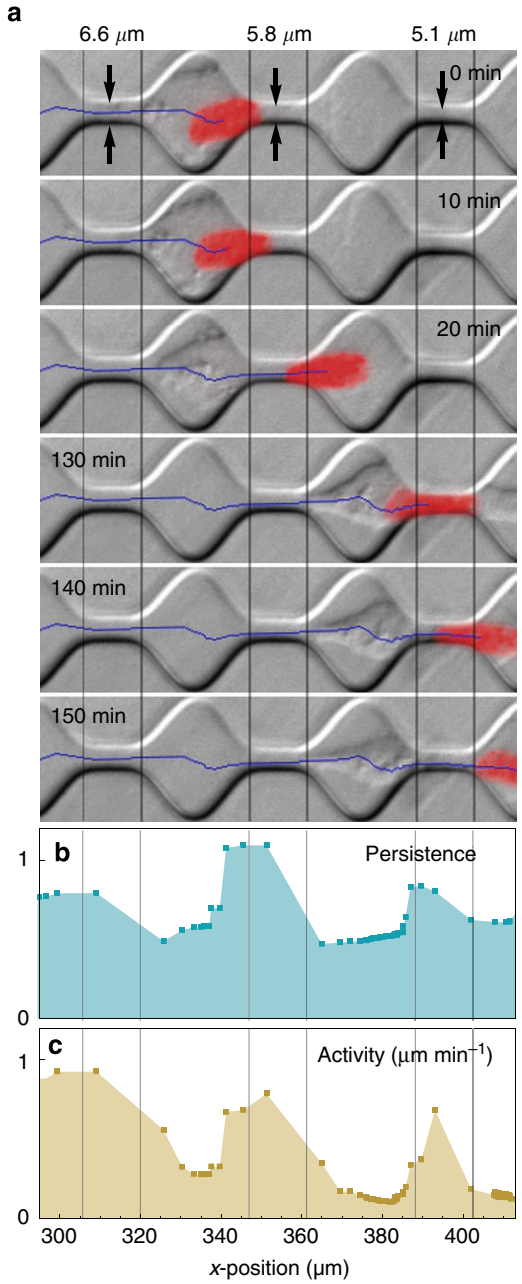


Figure 6 | Cell in a microstructured channel array. (**a**) Primary breast cancer cell migrating through a linear array of sequentially narrowing channels and wider chambers. The cell nucleus (red) is continuously tracked, with the centroid positions marked in blue. Persistence (**b**) and activity (**c**) are high when the nucleus transverses narrow channels, and decrease when the nucleus enters the wider chambers.

Note 1). In addition, the SWD becomes leptocurtic, but it does not show the exponential distribution found in the experiments. Finally, using an asymmetric bivariate normal distribution with positive correlations between q and a (Fig. 5b, dashed grey ellipse), the SWD, MSD and correlation functions match the measured data nearly perfectly (Fig. 5a,c,d, dashed grey line).

This example demonstrates how superstatistics can recapitulate the anomalous features of heterogeneous random walks by mapping the complexity of the system into a suitable distribution of parameter values $p_{\mathrm{model}}(q,a)$, while keeping the underlying stochastic process simple.

Discussion

In this study, we have applied the superstatistical framework to the specific example of tumour cell migration in environments with different dimensionality. The same approach, including the particular choice of the AR-1 process as a local model, can be used for many other heterogeneous random walks in life sciences. For this purpose, we provide a Python implementation of the Bayesian algorithm for inferring the time-dependent parameters q_t and a_t from random walk trajectories (Supplementary Software 1).

In principle, a sequential, grid-based inference of superstatistical parameters can also be performed by a Markov Chain Monte Carlo approach. In this case, the vector of model parameters to be inferred consists of the full set $\{(q_t,a_t)\}$ of superstatistical parameters for all time points. In the past, Markov Chain Monte Carlo methods, mostly based on the Metropolis Hastings algorithm, exhibited serious convergence problems when applied to such high-dimensional parameter spaces. Only recently, a novel sampling method based on Hamiltonian Monte Carlo has markedly improved the convergence[29]. Our preliminary tests demonstrate that this new sampling algorithm can indeed find the parameter vector of a hierarchical superstatistical model, however, with a considerably longer computation time.

Our superstatistical framework can be readily adapted to more complex types of stochastic systems. In particular, the AR-1 process can be replaced by any parameterized model with a defined likelihood function. For example, fluorescent beads attached to the cytoskeleton of living cells show fluctuations that can be described by a particle diffusing in a harmonic potential well[30,31]. Due to cytoskeletal remodelling, the centre position of the potential well is changing on longer timescales. Together, this process can be modelled with an inhomogeneous random walk of the centre position, superposed with a harmonic overdamped oscillator[32]. As a final example, recordings of neural spike trains are frequently modelled as inhomogeneous Poisson processes with a time-dependent spike rate. In this case, sequential Bayesian inference can be used to extract the local spike rates from the time series of measured interspike intervals.

Methods

Cell culture and migration measurements. For migration experiments in collagen, on plastic and on fibronectin-coated plastic, we use MDA-MB-231 breast carcinoma cells (obtained from the American Type Culture Collection (ATCC)). Cells are cultured in 75 cm^2 flasks in Dulbecco s modified Eagle's medium (DMEM) (1 $g\,l^{-1}$ D-glucose) and 10% fetal bovine serum, 1% penicillin/streptomycin at 37 °C, 5% CO_2 and 95% humidity. Cells are passaged every second day. Trypsin-ethylenediaminetetraacetic acid (Trypsin-EDTA) is used to detach cells.

To study cell migration on planar surfaces, we use tissue culture-treated plastic dishes with and without fibronectin coating (69 and 177 cells, respectively). In all 2D experiments, the sample time interval between frames was $\delta t_{2D} = 1$ min.

For 3D experiments, we use reconstituted collagen gels (Fig. 1a) with controlled material properties as a substitute for biological tissue. At a collagen concentration of 2.4 $mg\,ml^{-1}$, these gels have an average pore radius of 1.3 μm and a shear modulus of 108 Pa (ref. 33). Cells are mixed with collagen solution before polymerization at a concentration of 15,000 cells per ml. The x-, y- and z-position of the cells within the collagen gel is determined from a characteristic intensity profile of the refraction pattern around the nucleus of the cell (inset of Fig. 1b). A 3D deconvolution of the intensity profile then defines the cell position with an accuracy of 2 μm (r.m.s.). Cell tracking is performed automatically in real time, and the cell position is used to keep the motorized microscope x–y-centred and z-focused onto the cell at all times. Using a time-sharing mode, we are able to observe and follow up to 20 individual migrating cells within the same cell culture well over prolonged time periods (24 h). We record discrete cell positions with a sample time interval of $\delta t_{3D} = 2.5$ min (Fig. 1c). Cells undergoing cell division during the time of observation were excluded. The number of analysed cells in collagen was 65.

We also study the migration of primary inflammatory ductal breast cancer cells (gift from Pamela Strissel and Reiner Strick, Womens Hospital, University Clinics Erlangen) within a microfabricated channel structure made of polydimethylsiloxan. The structure has a constant height of 3.7 μm and consists of 15 consecutive channels with diameters decreasing from 11 to 1.7 μm, separated by 20 × 20-μm-

wide chambers (Fig. 6a). After staining the cell nuclei with Hoechst 33342 ($1\,\mu g\,ml^{-1}$), the centre positions are tracked with a sample time interval of $\delta t_{1D} = 5$ min. For superstatistical evaluation, a cell is chosen that passed through two successive channels within 150 min.

Bayesian parameter inference. Since the iterative updating of the parameter distribution described in this work is not analytically tractable, the presented algorithm is implemented using discretized probability distributions. Based on equally spaced parameter values q_i and a_j ($i \in \{1,2,..,N_q\}$, $j \in \{1,2,..,N_a\}$), a distribution $p(q,a)$ can be approximated by a $N_q \times N_a$-dimensional matrix: $(p(q,a))_{ij} = p(q = q_i, a = a_j)$. The multiplication of two distributions is thus reduced to the element-wise multiplication of two matrices.

The prior distribution $P_t = p(q_{t+1}, a_{t+1})$ holds the preliminary belief about the latent parameter values for the next time step, before seeing the corresponding data point. Using the data point $\mathbf{u}_{t+1}$, we subsequently update the prior distribution by multiplying it with the likelihood $L_{t+1} = p(\mathbf{u}_{t+1}|q_{t+1}, a_{t+1}; \mathbf{u}_t)$ that describes the probability of observing a certain measurement $\mathbf{u}_{t+1}$, given the values of the latent parameters (and the previous measurement $\mathbf{u}_t$).

For the underlying AR-1 process, the likelihood is given by

$$p(\mathbf{u}_{t+1}|q_{t+1}, a_{t+1}; \mathbf{u}_t) = \frac{1}{\left(2\pi a_{t+1}^2\right)^{d/2}} \exp\left(- \frac{(\mathbf{u}_{t+1} - q_{t+1}\mathbf{u}_t)^2}{2a_{t+1}^2} \right),$$

where d states the number of dimension of the velocity vectors (two in this study). Note that the inference method can also be applied to other underlying stochastic processes with more complicated likelihood functions. As our approach uses only the numerical values of the likelihood for discrete points of the (q_t,a_t)-grid, the likelihood need not be expressed analytically as long as it can be computed numerically.

The next prior P_{t+1} is computed from the posterior distribution $P_{t+1} = K(P_t L_{t+1})$, with K being a transformation that accounts for both gradual and abrupt parameter changes as follows: To allow for abrupt parameter changes, we set the minimal probability of the posterior distribution to $p_{\min} = 10^{-7}$

$$P_t L_{t+1} \rightarrow \max[p_{\min}, P_t L_{t+1}].$$

To allow for gradual parameter changes, we blur the distribution by convolution with a box kernel B of radius $R = 0.03$ defined as

$$B(q,a) = \Theta(R - |q|) \cdot \Theta(R - |a|).$$

Here, $\Theta(x)$ is the Heaviside step function. The posterior distribution of the parameters is normalized at every time step, since the transformation K does not preserve normalization. A systematic procedure to find optimal values for the two parameters $p_{\min}$ and R is given in the Supplementary Note 5.

Starting with a flat prior P_0 and moving forward in time using the iteration described above, a series of 'forward' priors $\{P_t^F\}_t$ is generated. In the same way, we can start the iteration at the end of a trajectory, and build a series of 'backward' prior distributions $\{P_t^B\}_t$. Finally, for each time step t, we multiply the $t-1$ and $t+1$ priors with the likelihood L_t to compute the final posterior distribution of the parameters (q_t,a_t), so that $P_{t-1}^F \; P_{t+1}^B \; L_t$. Note that the inference algorithm is run in both directions of time to ensure that for each estimated parameter pair (q_t,a_t), all measured data points are taken into account and not only those of earlier times $0 \ldots t$. In principle, however, the algorithm can also be used only in the forward direction, which may be useful for online analysis of a data stream.

Temporal and ensemble averages. Throughout this paper, the symbol $\langle f_t \rangle_t$ denotes temporal averaging over all discrete time points. For our data evaluation (SWD, MSD and auto- and cross-correlations), we have additionally ensemble-averaged the time-averaged properties over the individual cells of the same migration environment.

Auto- and cross-correlations. The auto-correlation $C_{qq}(\Delta t)$ of the persistence parameter q_t is defined in the standard way as $C_{qq}(\Delta t) = \frac{\langle (q_t - \bar{q})(q_{t+\Delta t} - \bar{q}) \rangle_t}{\sigma_q^2}$, where $\bar{q} = \langle q_t \rangle_t$ is the temporal average and $\sigma_q^2 = \left\langle (q_t - \bar{q})^2 \right\rangle_t$ is the variance of the parameter. The definition of the activity auto-correlation $C_{aa}(\Delta t)$ is analogous. Finally, the cross-correlation $C_{qa}(\Delta t)$ between the two parameters is defined as $C_{qq}(\Delta t) = \frac{\langle (q_t - \bar{q})(a_{t+\Delta t} - \bar{a}) \rangle_t}{\sqrt{\sigma_q^2}\sqrt{\sigma_a^2}}$.

Superstatistical modelling of cell migration. To model the statistical properties of cell trajectories in collagen (Fig. 5, grey dashed lines), we use a superstatistical regime-switching process with an average switching time of $\tau = 200$ min. Parameter values (q_t,a_t) are drawn from a bivariate Gaussian distribution, $(q_t,a_t) \sim N(\mu,\Sigma)$, centred around the mean $\boldsymbol{\mu} = (\mu_q,\mu_a) = (-0.05, 0.55)$. The covariance matrix is $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$ with $\sigma = 0.3$ and $\rho = 0.65$. The 50% credibility region of the distribution is shown in Fig. 5b as a grey dashed ellipse. The values of q_t are restricted to the interval $[-1,1]$.

Environment-specific cell classification. For 'leave-one-out' cross-validation, we calculate the squared deviation D between the time-averaged posterior distribution of a single cell, denoted $p_{\text{single}}(q,a)$, and each of the three ensemble- and time-averaged distributions $p_{\text{env}}(q,a)$ (excluding that one cell). The calculation of the squared deviation is carried out as a sum over the $N_q \times N_a$-grid:

$$D = \sum_{i=1}^{N_q} \sum_{j=1}^{N_a} \left(p_{\text{single}}(q = q_i, a = a_j) - p_{\text{env}}(q = q_i, a = a_j) \right)^2$$

A cell is counted as correctly classified if the deviation to its true environment is the smallest, compared with the other two environments.

References

1. Pedrycz, W. & Chen, S. (eds.) *Time Series Analysis, Modeling and Applications* (Springer, 2013).
2. Zumbach, G. *Discrete Time Series, Processes, and Applications in Finance* (Springer, 2013).
3. Kirchgässner, G., Wolters, J. & Hassler, U. *Introduction to Modern Time Series Analysis* (Springer, 2013).
4. Rabiner, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–286 (1989).
5. Wu, P.-H., Giri, A., Sun, S. X. & Wirtz, D. Three-dimensional cell migration does not follow a random walk. *Proc. Natl Acad. Sci. USA* **111,** 3949–3954 (2014).
6. Bursac, P. *et al.* Cytoskeletal remodelling and slow dynamics in the living cell. *Nat. Mater.* **4,** 557–561 (2005).
7. Dieterich, P., Klages, R., Preuss, R. & Schwab, A. Anomalous dynamics of cell migration. *Proc. Natl Acad. Sci. USA* **105,** 459–463 (2008).
8. Beck, C. & Cohen, E. Superstatistics. *Physica A* **322,** 267–275 (2003).
9. Beck, C., Cohen, E. & Swinney, H. From time series to superstatistics. *Phys. Rev. E* **72 (2005).**
10. Beck, C. Generalized statistical mechanics for superstatistical systems *Phil. Trans. R. Soc. A* **369,** 453–465 (2011).
11. Van der Straeten, E. & Beck, C. Superstatistical fluctuations in time series: Applications to share-price dynamics and turbulence. *Phys. Rev. E* **80,** 036108 (2009).
12. Rorth, P. Collective cell migration. *Annu. Rev. Cell. Dev. Biol.* **25,** 407–429 (2009).
13. Rorth, P. Fellow travellers: emergent properties of collective cell migration. *EMBO Rep.* **13,** 984–991 (2012).
14. Friedl, P. & Gilmour, D. Collective cell migration in morphogenesis, regeneration and cancer. *Nat. Rev. Mol. Cell Biol.* **10,** 445–457 (2009).
15. Potdar, A. A., Jeon, J., Weaver, A. M., Quaranta, V. & Cummings, P. T. Human mammary epithelial cells exhibit a bimodal correlated random walk pattern. *PLoS ONE* **5,** e9636 (2010).
16. Demou, Z. N. & McIntire, L. V. Fully automated three-dimensional tracking of cancer cells in collagen gels: determination of motility phenotypes at the cellular level. *Cancer Res.* **62,** 5301–5307 (2002).
17. Niggemann, B. *et al.* Tumor cell locomotion: differential dynamics of spontaneous and induced migration in a 3d collagen matrix. *Exp. Cell Res.* **298,** 178–187 (2004).
18. Takagi, H., Sato, M. J., Yanagida, T. & Ueda, M. Functional analysis of spontaneous cell movement under different physiological conditions. *PLoS ONE* **3,** e2648 (2008).
19. Selmeczi, D. *et al.* Cell motility as random motion: A review. *Eur. Phys. J.* **157,** 1–15 (2008).
20. Bischofs, I. B. & Schwarz, U. S. Cell organization in soft media due to active mechanosensing. *Proc. Natl Acad. Sci. USA* **100,** 9274–9279 (2003).
21. Koch, T. M., Münster, S., Bonakdar, N., Butler, J. P. & Fabry, B. 3d traction forces in cancer cell invasion. *PLoS ONE* **7,** e33476 (2012).
22. Höfling, F., Frey, E. & Franosch, T. Enhanced diffusion of a needle in a planar array of point obstacles. *Phys. Rev. Lett.* **101,** 120605 (2008).
23. Dickinson, R. B., Guido, S. & Tranquillo, R. T. Biased cell migration of fibroblasts exhibiting contact guidance in oriented collagen gels. *Ann. Biomed. Eng.* **22,** 342–356 (1994).
24. Provenzano, P. P., Inman, D. R., Eliceiri, K. W., Trier, S. M. & Keely, P. J. Contact guidance mediated three-dimensional cell migration is regulated by rho/rock-dependent matrix reorganization. *Biophys. J.* **95,** 5374–5384 (2008).
25. DiMilla, P. A., Stone, J. A., Quinn, J. A., Albelda, S. M. & Lauffenburger, D. A. Maximal migration of human smooth muscle cells on fibronectin and type iv collagen occurs at an intermediate attachment strength. *J. Cell Biol.* **122,** 729–737 (1993).
26. Maiuri, P. *et al.* Actin flows mediate a universal coupling between theory actin flows mediate a universal coupling between cell speed and cell persistence. *Cell* 1–13 (2015).
27. Mierke, C. T., Frey, B., Fellner, M., Herrmann, M. & Fabry, B. Integrin $\alpha 5\beta 1$ facilitates cancer cell invasion through enhanced contractile forces. *J. Cell. Sci.* **124,** 369–383 (2011).

28. Hawkins, R. J. *et al.* Pushing off the walls: a mechanism of cell motility in confinement. *Phys. Rev. Lett.* **102,** 1–4 (2009).
29. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15,** 1351–1381 (2014).
30. Metzner, C., Raupach, C., Paranhos Zitterbart, D. & Fabry, B. Simple model of cytoskeletal fluctuations. *Phys. Rev. E* **76,** 021925 (2007).
31. Raupach, C. *et al.* Stress fluctuations and motion of cytoskeletal-bound markers. *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.* **76,** 011918–011918 (2007).
32. Metzner, C., Raupach, C., Mierke, C. T. & Fabry, B. Fluctuations of cytoskeleton-bound microbeads—the effect of bead–receptor binding dynamics. *J. Phys. Condens. Matter* **22,** 194105 (2010).
33. Mickel, W. *et al.* Robust pore size analysis of filamentous networks from three-dimensional confocal microscopy. *Biophys. J.* **95,** 6072–6080 (2008).

Acknowledgements

We thank Thorsten Koch for help with image acquisition, and Caroline Gluth and Amy Rowat for helping with microfluidic device design and preparation. We also thank Pamela Strissel and Reiner Strick (University of Erlangen, University Clinics) for establishment of a primary inflammatory breast cancer cell line and for sharing these cells with our laboratory. This work was supported by the Deutsche Forschungsgemeinschaft, the Research Training Group 1962 'Dynamic Interactions at Biological Membranes: From Single Molecules to Tissue', and the National Institutes of Health.

Author contributions

C.Me. and B.F. designed the study. J.S. and L.L. developed the data acquisition software and performed the cell experiments. C.Me., C.Ma. and F.S. developed the theoretical model and analyzed the data. C.Me., B.F. and C.Ma. wrote the paper. All authors read and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Metzner, C. *et al.* Superstatistical analysis and modelling of heterogeneous random walks. *Nat. Commun.* 6:7516 doi: 10.1038/ncomms8516 (2015).

Supplementary Note 1

Emergence of anomalous features in time-averaging statistics of heterogeneous random walks

The traditional tools for analyzing random walks are step width distributions (SWD) and mean squared displacements (MSD). For a given trajectory time series $\left\{\vec{R}_t^{(s)}\right\} = \left\{(x_t^{(s)}, y_t^{(s)}, z_t^{(s)})\right\}$, these statistical measures are computed by averaging certain quantities of interest over the time index t. An additional average may be performed over an ensemble of different measured time series s. In particular, the SWD in x-direction is defined as

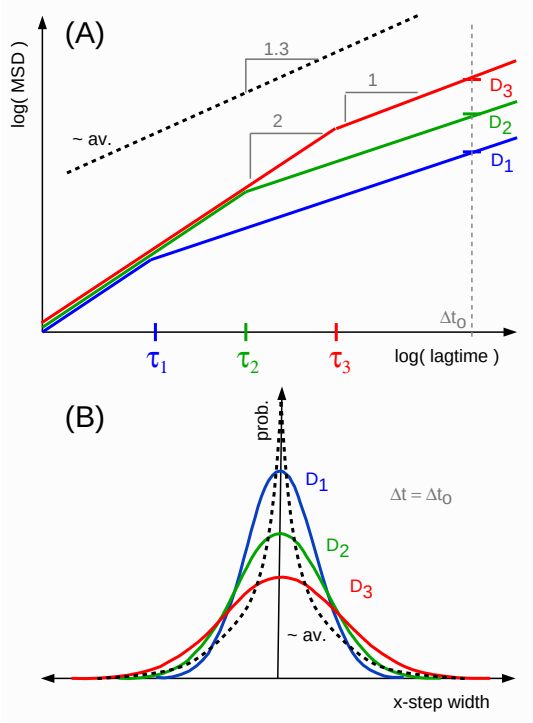
$$P(\Delta x, \Delta t) = \left\langle \delta\left(\Delta x - (x_{t+\Delta t}^{(s)} - x_t^{(s)})\right)\right\rangle_{t,s} \tag{1}$$

and the MSD is defined as

$$\overline{R^2}(\Delta t) = \left\langle |\vec{R}_{t+\Delta t}^{(s)} - \vec{R}_t^{(s)}|^2 \right\rangle_{t,s}. \tag{2}$$

In the case of an uncorrelated random walk, the SWD is a Gaussian distribution with respect to Δx, with a variance that increases linearly as a function of the lag-time Δt. The MSD also increases linearly for all lag-times.

For a homogeneous correlated random walk with correlation time τ, the behavior of the SWD and MSD depend on the lag-time Δt. In the regime of short lag-times, $\Delta t \ll \tau$, the SWD is non-Gaussian and the MSD grows quadratically (ballistic motion). In the regime of long lag-times, $\Delta t \gg \tau$, the SWD is Gaussian and the MSD grows linearly (diffusive motion), just as in an uncorrelated random walk. The correlation time τ thus defines the cross-over time between the slope 2 and slope 1 in a double-logarithmic plot of the MSD versus lag-time (colored lines in Supplementary Figure 1A).



Supplementary Figure 1: ***Emergence of anomalous mean squared displacements and step width distributions***. *Anomalous statictics can be due to superstatistical fluctuations of the velocity decorrelation time τ. (A): Mean squared displacements (MSD, colored lines) for three different decorrelation times τ_s, which mark the cross-over points from the ballistic regime to the diffusive regime. Averaging over an ensemble with suitably distributed τ_s can result in a power law with fractional exponent (black dashed line). For lag-times $\Delta t \gg \tau_s$, each case s can be described by an effective diffusivity D_s. (B): Step width distributions (SWD, colored lines) at a sufficiently long lag-time Δt_0 are Gaussians with a variance proportional to D_s. Averaging over an ensemble of narrow and wide normal distributions results in a leptocurtic distribution (black dashed line).*

Next, we consider a heterogeneous ensemble $\left\{\vec{R}_t^{(s)}\right\}$ of temporally homogeneous correlated random walks, each with a different correlation time τ_s. In this case, the SWD in the regime of long lag-times is an average over Gaussians of different variance (colored curves in Supplementary Figure 1B), which results in a leptocurtic and therefore anomalous distribution (black dashed curve in Supplementary Figure 1B). Similarly, for a sufficiently wide and dense distribution of τ_s, the individual cross-over times between ballistic (slope 2) and diffusive (slope 1) regimes average out, and the MSD assumes an anomalous shape. In particular, it can be shown that heterogeneity of this kind can lead to an almost exponential SWD and a MSD that resembles a power law with a fractional exponent (dashed line in Supplementary Figure 1A). For an analytical derivation, see arXiv:1207.2242).

Finally, note that a temporally heterogeneous random walk in which the correlation time τ varies slowly as a function of time t has a similar effect as the heterogeneous ensemble above. This demonstrates that anomalous features emerge naturally in the SWD and MSD if these statistics are applied to heterogeneous random walks.

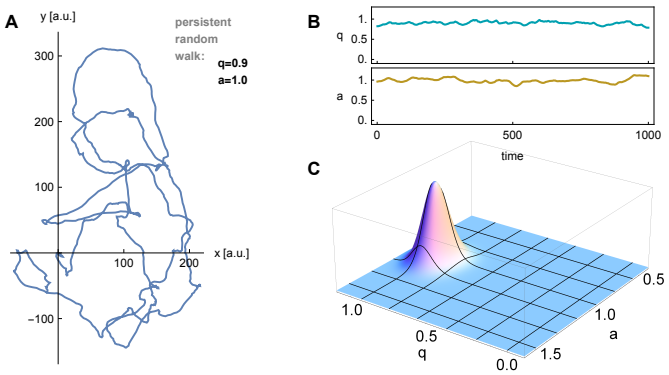
Supplementary Note 2

Random walk models in the superstatistical framework

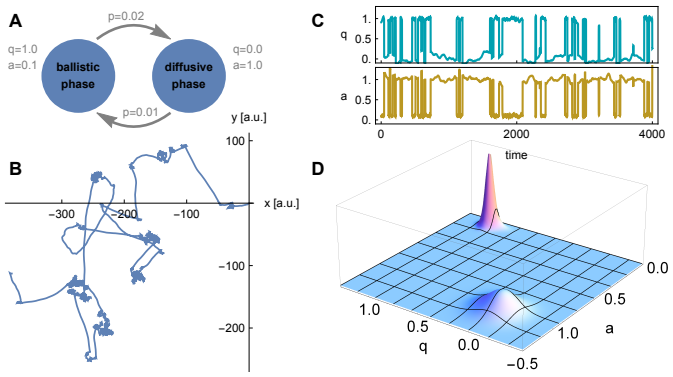
In a superstatistical framework, random time series are described, on the lowest hierarchical level, by a simple stochastic 'base model' with a small set of parameters. In this paper, we have chosen the AR-1 process with a persistence parameter q and an activity parameter a. Keeping these superstatistical model parameters constant corresponds to a temporally homogeneous random process. For example, the AR-1 process with constant parameters q and a corresponds to a homogeneous persistent random walk (Supplementary Figure 2).

In order to describe more complex time series, in particular measured random walks that show obvious temporal heterogeneity, it is possible to resort to higher order base models, such as the class of AR-n processes with $n>1$. However, this approach is misguided in cases where the temporal heterogeneity is caused by time- or position-dependent external influences that cannot be easily incorporated into the base model. A standard example of a random walk that becomes heterogeneous due to external effects is a particle diffusing in a medium with different local temperatures.

The superstatistical framework acknowledges external time-dependent effects and models them directly as a higher level stochastic process that affects the parameters of the base model. In our case, the two parameters q_t and a_t of the AR-1 process are allowed to change at any time step, according to some higher order process. By this way, heterogeneous random walks of arbitrary complexity can be described by a hierarchy of low-order stochastic models. A simple example that highlights externally controlled heterogeneity is the intermittent random walk, in which a particle switches randomly between phases of ballistic and diffusive motion. This heterogeneous process can be described by a two-level superstatistical model with an AR-1 process as the base model and a Markov switching process as the second level model (Supplementary Figure 3).



Supplementary Figure 2: ***Persistent random walk in 2D***. *Simulated random walk, generated by using an AR-1 process with constant values $q = 0.9$ for the persistence and $a = 1$ for the activity. (A): Sample trajectory. (B): Superstatistical parameters q_t and a_t as a function of time, reconstructed from the sample trajectory. (C): Corresponding time-averaged posterior distribution $p(q, a)$. The single peak indicates that the random walk is temporally homogeneous.*

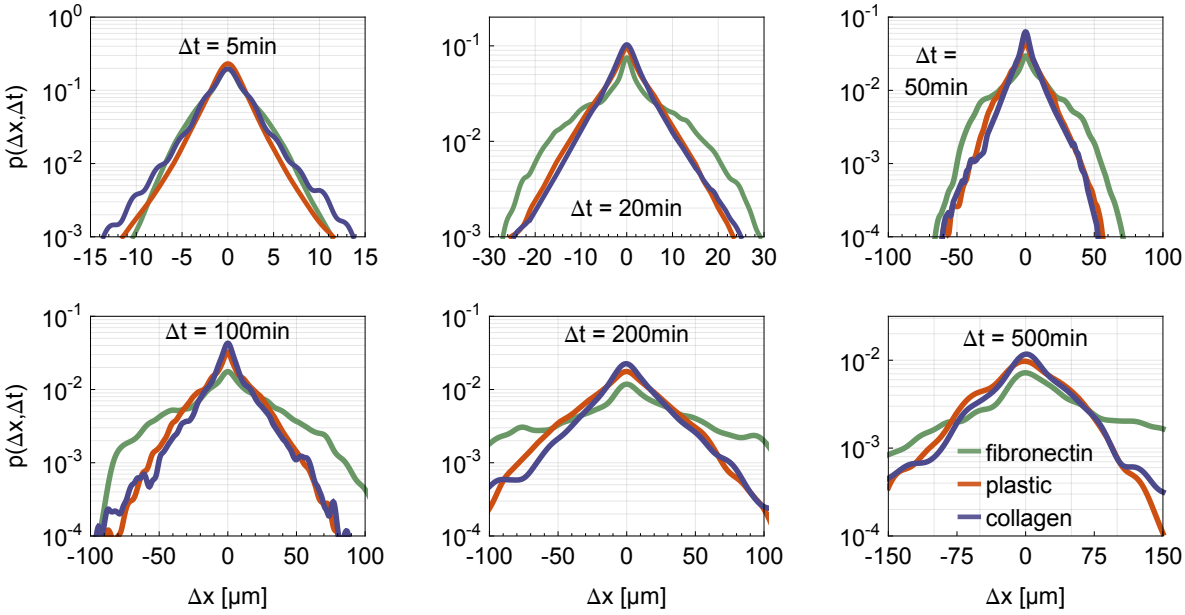


Supplementary Figure 3: ***Intermittent random walk in 2D***. *Simulated random walk, generated by an AR-1 process in which the superstatistical parameters switch stochastically between a ballistic phase* $(q = 1, a = 0.1)$ *and a diffusive phase* $(q = 0, a = 1.0)$. *(A) Representation of the Markov process that controls the stochastic switching between the two phases. (B) Sample trajectory. (C): Superstatistical parameters* q_t *and* a_t *as a function of time, reconstructed from the sample trajectory. (D): Corresponding time-averaged posterior distribution* $p(q, a)$. *The two peaks directly reflect the two distinct phases in this temporally heterogeneous random walk.*

Supplementary Note 3

Step width distributions

In the main part of the paper, we have argued that the traditional, globally averaging statistical measures, such as the mean squared displacements (MSD) and the step width distributions (SWD), are often masking important differences between distinct types of random walks. In particular, we have shown that the SWD of MDA-MB-231 breast carcinoma cells migrating in a collagen network and on a plastic dish are practically indistinguishable on a time scale of 120 min. Here, we demonstrate that the SWDs in the two environments are also nearly identical for all lag times between 5 min and 500 min (Supplementary Figure 4, blue and red lines). The form of the distributions remains approximately exponential for the whole range of lag-times. In contrast, the SWD for migration on fibronectin coated surfaces (green lines) is significantly different from the other two environments for all lag times larger than 10 min.

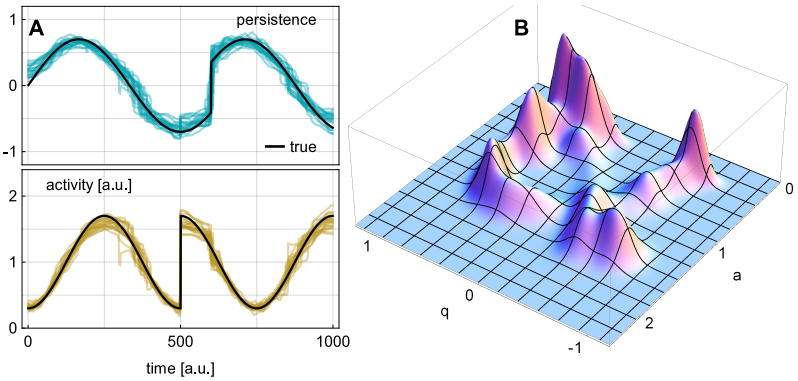


Supplementary Figure 4: ***Step width distributions (SWDs)***. *Shown are SWDs (for six different lag-times) of MDA-MB-231 breast carcinoma cells in a collagen network (blue), on uncoated plastic (red) and on plastic coated with the adhesive ligand fibronectin. The distributions in collagen and on plastic are similar for all lag-times, although the migration mechanism of the cells is drastically different in these two environments.*

Supplementary Note 4

Comparison of Bayesian parameter inference with sliding window-based Maximum Likelihood estimation

In the main text, we have shown that the presented Bayesian method yields accurate results for simulated random-walk trajectories with abruptly switching (Fig. 3a,b) as well as linearly changing parameter values (Fig. 3c,d). Here, we provide an additional test case by combining both abrupt and gradual, sinusoidal parameter changes (Supplementary Figure 5). Based on these three test cases (regime-switching, linear, and sinusoidal), we compare our Bayesian method to an alternative inference method for heterogeneous random walks, namely the Maximum Likelihood estimation using a sliding window.



Supplementary Figure 5: ***Evaluation of the Bayesian inference method.*** *We use simulated trajectories with abrupt as well as gradual, sinusoidal parameter variations. (A) The reconstructed posterior mean values of persistence and activity from 20 simulated trajectories (color) are plotted together with the true values (black). (B) The time-averaged posterior distribution corresponding to a single trajectory reveals the complex correlations between the two parameters, persistence and activity.*

Sliding window analysis

The standard method for inferring time-dependent parameter changes in time series is the sliding window method, whereby the time series is subdivided into overlapping segments. For each segment, the parameters (activity and persistence in our case) are assumed to be constant and can be estimated using a maximum likelihood approach. Considering a homogeneous AR(1)-process for times t' within the interval $I_t = \{t - w/2, ..., t + w/2\}$ of length w, centered around some point in time t, with persistence q_t and activity a_t, one can state the log-likelihood for these parameters as follows:

$$\log\ p\left(\{\vec{u}_{t'}\}_{t'\in I_t}|q_t, a_t\right) \ \propto \sum_{t'\in I_t}\left(-\frac{\left(\vec{u}_{t'} - q_t\vec{u}_{t'-1}\right)^2}{2a_t^2} - \log\left(2\pi a_t^2\right)\right) \ .$$

Here, we consider a two-dimensional trajectory. Maximizing the log-likelihood with respect to q_t and a_t yields the following Maximum Likelihood estimators:

$$\hat{q}_t = \frac{\sum_{t'\in I_t} \vec{u}_{t'}.\vec{u}_{t'-1}}{\sum_{t'\in I_t} \vec{u}_{t'-1}.\vec{u}_{t'-1}} \quad \text{and} \quad \hat{a}_t = \sqrt{\frac{1}{2w}\sum_{t'\in I_t}\left(\vec{u}_{t'} - \hat{q}_t\vec{u}_{t'-1}\right)^2} \ ,$$

where (.) denotes the dot product. Note that the estimator of a_t depends on the estimated value of q_t. To analyze a heterogeneous time series using these estimators, the time series is partitioned into overlapping segments of length w – thus the term 'sliding window'. This segmentation allows us to estimate the local persistence and activity over the entire trajectory.

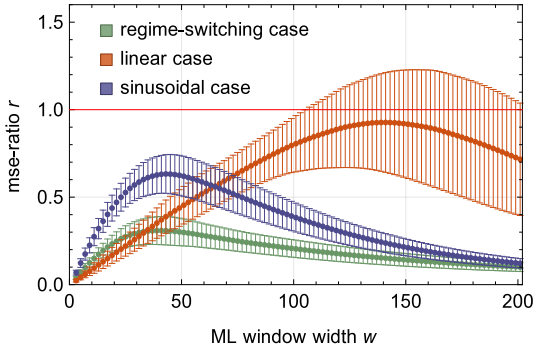
Method comparison

We use the mean squared error (mse) of the estimated parameter sequences with respect to the true parameter sequences to assess the quality of the two methods:

$$\mathrm{mse}\left(\{(q_t, a_t), (\hat{q}_t, \hat{a}_t)\}\right) = \frac{1}{2N'} \sum_t \left((q_t - \hat{q}_t)^2 + (a_t - \hat{a}_t)^2 \right) \ ,$$

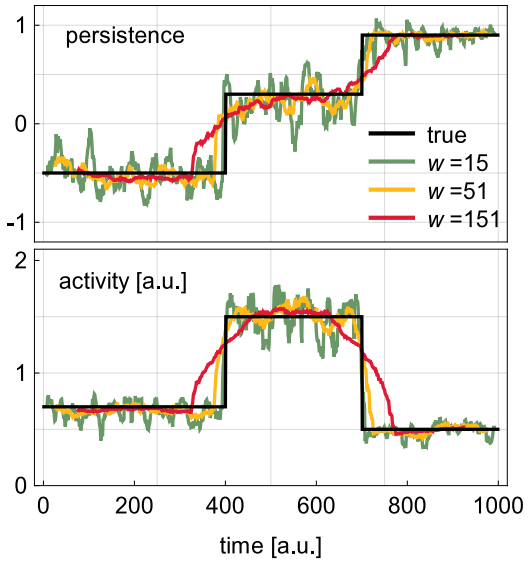
where N' denotes the length of the estimated parameter sequence. The ratio of the errors between both methods allows for a direct comparison:

$$r(w) = \frac{\mathrm{mse}\left(\{(q_t, a_t), (\hat{q}_t, \hat{a}_t)_{\mathrm{Bayes}}\}\right)}{\mathrm{mse}\left(\{(q_t, a_t), (\hat{q}_t, \hat{a}_t)_{\mathrm{ML}(w)}\}\right)} \ .$$



Supplementary Figure 6: ***Comparison of Bayesian inference with the Maximum Likelihood (ML) approach***. *Based on the mean squared error ratio, the Bayesian method is directly compared to the ML approach for a variety of window sizes. With an average mse ratio smaller than one for all three test cases, the Bayesian method is found to be superior, regardless of the window size for the ML approach.*

We estimate r for different window widths ranging from 3 to 200 data points. For all three test cases (regime-switching, linear, and sinusoidal), and regardless of the chosen width of the sliding window, the Bayesian algorithm attains a smaller error on average ($r < 1$, see Supplementary Figure 6). Apart from the smaller estimation error, the Bayesian approach provides an advantage especially for small data sets. For the sliding window analysis, no parameters can be estimated during a time period of $w/2$ at the beginning and the end of the time series, due to the finite width of the sliding window. By contrast, the Bayesian method yields $N - 1$ estimates for N data points. Furthermore, the assumption of constant parameters q and a within each window of length w is an inherent limitation of the sliding window approach, since it is violated for most heterogeneous time series. Finally, the choice of the window size w strongly affects the resulting reconstruction of the time-varying parameters. For small window sizes, the Maximum Likelihood method can detect sudden parameter changes but shows large, erroneous fluctuations in the estimated parameter sequence (Supplementary Figure 7, green line). With increasing window size, these erroneous fluctuations diminish, but sudden parameter changes become increasingly blurred (Supplementary Figure 7, yellow and red).



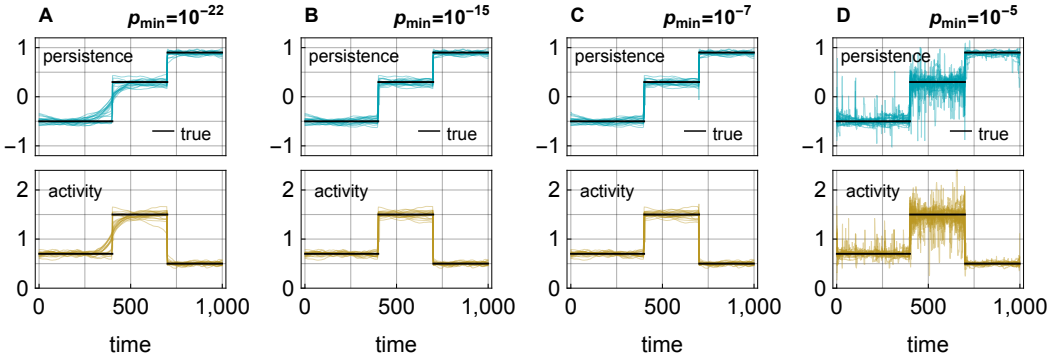
Supplementary Figure 7: ***Influence of window size***. *Effect of window size on Maximum Likelihood estimation in a regime-switching case (see Fig. 3). The true parameter values (black) are shown together with the estimated persistence and activity for different window sizes (color).*

Supplementary Note 5

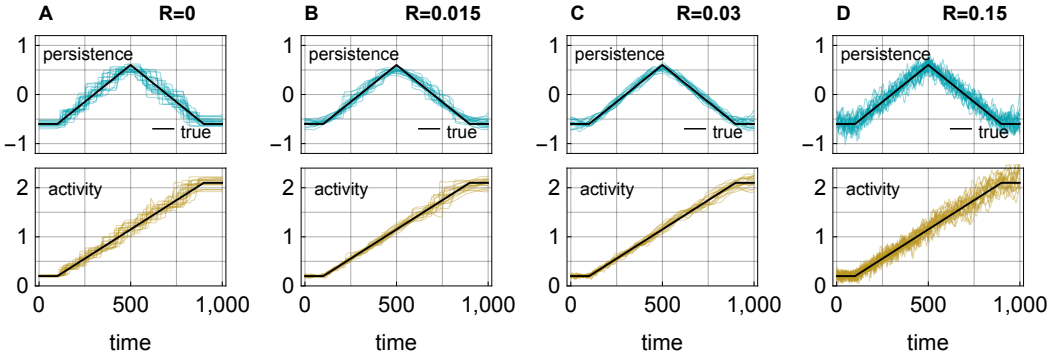
Choice of the kernel parameters p_{min} and R.

In the main text, we have presented a sequential Bayesian method to reconstruct the superstatistical parameters q_t and a_t from a measured random walk time series. In each time step, the posterior distribution $P_t(q, a)$ is given a minimum probability $p_{min} = 10^{-7}$, which accounts for the possibility of abrupt parameter jumps to any value within the grid. Furthermore, the posterior distribution is convolved with a box kernel of radius $R = 0.03$, which assigns small (gradual) parameter changes a larger probability compared to large jumps.

In general, the two control parameters p_{min} and R do not require a precise fine tuning, but setting them outside a certain viable range leads to predictable problems: In general, too small values for the control parameters lead to a blurring of rapid parameter changes, whereas too large values lead to overfitting. We illustrate the effects of wrongly chosen control parameters p_{min} and R in Supplementary Figures 8 and 9.



Supplementary Figure 8: ***Effect of kernel parameters for sudden changes of persistence and activity.*** *Effects of a too small (A), about optimal (B,C) and too large (D) setting of the control parameter p_{min} on the reconstruction of abruptly jumping superstatistical parameters q_t (top) and a_t (bottom). Note that the reconstruction quality degrades only very weakly, even if the control parameter deviates from the optimum value by orders of magnitude.*



Supplementary Figure 9: ***Effect of kernel parameters for gradual changes of persistence and activity.*** *Effects of a too small (A), about optimal (B,C) and too large (D) setting of the control parameter R on the reconstruction of more gradually evolving superstatistical parameters q_t (top) and a_t (bottom).*

Optimum values for the two control parameters p_{min} and R can be found systematically by generating synthetic data sets that resemble the actual data and by quantifying how well the known superstatistical parameter changes of (q_t, a_t) are reconstructed by the method. The values for p_{min} and R given in the paper have been obtained in this way.

If the dynamics of the superstatistical parameters (q_t, a_t) is entirely unknown, the generation of realistic synthetic data sets is impossible. In such cases, the choice of the control parameters p_{min} and R can be viewed as a Bayesian

model selection problem. Using measured data only, the plausibility of a given set (p_{min}, R) can then be quantified by the global marginal likelihood L: For each time step t, the momentary marginal likelihood L_t is obtained by integrating the likelihood function over the (q_t, a_t)-plane. This integral can be evaluated readily within our grid based approach. The global marginal likelihood L is then given by the product of all L_t. The optimal set of control parameters, given the measured data, can then be found by maximizing L.

Supplementary Note 6

Persistence time and cell speed

In a superstatistical data analysis, the parameters q_t and a_t used in this work can be replaced by different variables that may be more appropriate for specific questions. For example, persistence can also be quantified by a persistence time τ_t, defined as the half decay time of the velocity correlations. Assuming $0 < q_t < 1$, and measuring the correlation time in units of the sampling time, τ_t is monotonically increasing with q_t, according to

$$\tau_t = \frac{\log(1/2)}{\log(q_t)}.$$

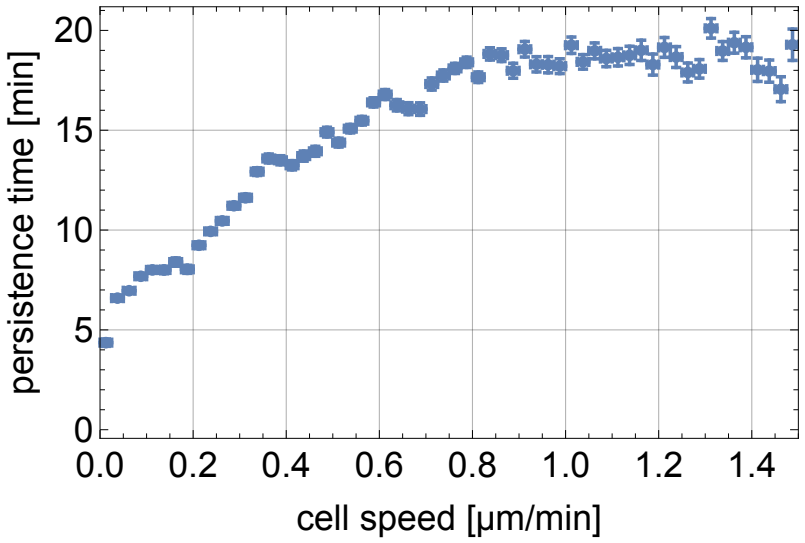
Furthermore, rather than investigating superstatistical parameter correlations between q_t and a_t, new insights can also be gained from the 'inter-level' correlations between superstatistical parameters and directly measured variables, such as the momentary cell speed u_t.

The relationship between cell speed, activity and cell persistence can be seen most easily in the case of a homogeneous AR-1 process in 2D. Assuming $0 \leq q_t < 1$, the momentary speed u_t is then Rayleigh distributed with the most probable value $\hat{u}_t$ being given by

$$\hat{u}_t = \frac{a_t}{\sqrt{1 - q_t^2}}.$$

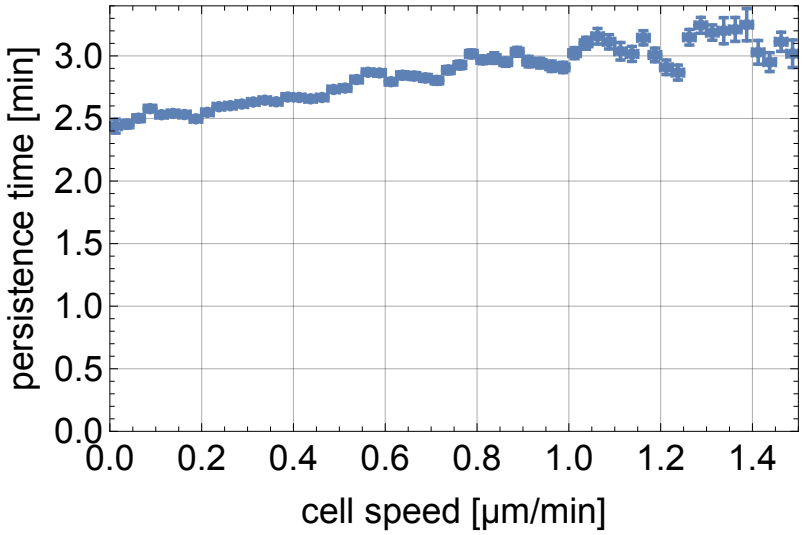
In a hypothetical situation with $a = constant$, there is always a positive correlation between $\hat{u}_t$ and q_t (or, equivalently, between u_t and τ_t). However, in the case of migrating cells, the situation is more complicated as the activity a_t is also a strongly time-varying quantitiy that may be correlated positively or negatively with the persistence q_t.

For example, Miauri et al. have recently reported an increase of the momentary persistence time τ_t with the momentary cell speed u_t for a variety of different cell types migrating on fibronectin-coated surfaces (Ref.[26] in the main manuscript). From our cell migration experiements on fibronectin-coated plastic, we obtain a similar dependence of τ_t on u_t, including the saturation of persistence time for larger cell speeds (Supplementary Figure 10).



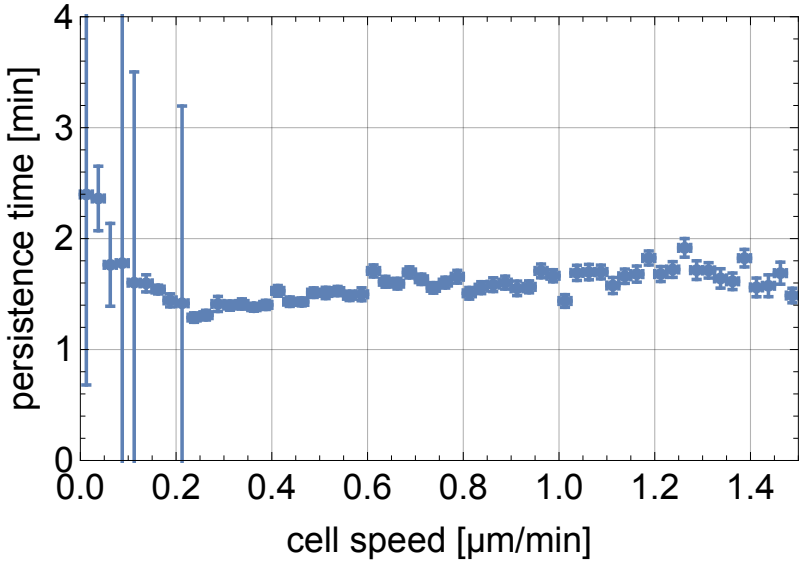
Supplementary Figure 10: ***Persistence time versus momentary speed for MDA-MB-231 cells on fibronectin-coated plastic.*** *Horizontal bars indicate the bins of cell speed, vertical bars the standard error of the mean persistence time.*

On uncoated plastic, where persistence is considerably lower, τ_t also increases with u_t, but without clear signs of saturation (Supplementary Figure 11).



Supplementary Figure 11: ***Persistence time versus momentary speed for MDA-MB-231 cells on uncoated plastic.*** *Horizontal bars indicate the bins of cell speed, vertical bars the standard error of the mean persistence time.*

For cell migration in collagen, we find only a weak dependence of τ_t on u_t (Supplementary Figure 12).



Supplementary Figure 12: ***Persistence time versus momentary speed for MDA-MB-231 cells in collagen.*** *Horizontal bars indicate the bins of cell speed, vertical bars the standard error of the mean persistence time.*