# Arbitrary Boolean logical search operations on massive molecular file systems

James L. Banal[1†], Tyson R. Shepherd[1†], Joseph Berleant[1†], Hellen Huang[1], Miguel Reyes[1,2], Cheri M. Ackerman[2], Paul C. Blainey[1,2,3], and Mark Bathe[1,2*]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA.

[3]Koch Institute for Integrative Cancer Research at MIT, Cambridge, MA 02142 USA.

[*]Correspondence should be addressed to: mark.bathe@mit.edu

[†]These authors contributed equally to this work.

**DNA is an ultra-high-density storage medium that could meet exponentially growing worldwide data storage demand. However, accessing arbitrary data subsets within exabyte-scale DNA data pools is limited by the finite addressing space for individual DNA-based blocks of data. Here, we form files by encapsulating data-encoding DNA within silica capsules that are surface-labeled with multiple unique barcodes. Barcoding is performed with single-stranded DNA representing file metadata that enables Boolean logic selection on the entire pool of data. We demonstrate encapsulation and Boolean selection of sub-pools of image files using fluorescence-activated sorting, with selection sensitivity of 1 in $10^6$ files per channel. Our strategy in principle enables retrieval of targeted data subsets from exabyte- and larger-scale data pools, thereby offering a random access file system for massive molecular data sets.**

DNA is the polymer used for storage and transmission of genetic information in biology. In principle, DNA can also be used as a medium for the storage of arbitrary digital information at densities far exceeding existing commercial data storage technologies and at scales well beyond the capacity of current data centers [1]. Ongoing advances in nucleic acid synthesis and sequencing technologies also continue to reduce dramatically the cost of writing and reading DNA, thereby rendering DNA-based digital information storage potentially viable economically in the near future [2-5]. As demonstrations of its viability as a general information storage medium, to date books, images, computer programs, audio clips, works of art, and Shakespeare's sonnets have all been stored in DNA using a variety of encoding schemes [6-12]. In each case, digital information was converted to DNA sequences and typically fragmented into 100–200 nucleotide (nt) blocks of data

for ease of chemical synthesis and sequencing. Sequence fragments were then assembled to reconstruct the original, encoded information.

While significant research effort has focused on improving DNA synthesis and encoding schemes, an additional, crucial aspect of digital data storage and retrieval is the ability to access specific subsets of a data pool on demand, which is conventionally achieved using polymerase chain reaction (PCR) [8,10,12]. PCR-based strategies take advantage of the ease of replication of DNA to extract specific DNA sequences from a DNA data pool using custom-designed forward and reverse primers that are complementary to the flanking sequences of interest. Nested addressing barcodes [13-15] can also be used to uniquely identify files using multiple barcodes. For an exabyte-scale data pool, each file requires at least four barcodes, or up to one hundred nucleotides in total barcode sequence length, thereby nearly eliminating the number of nucleotides that can be used for data encoding. Further, orthogonality of barcodes to other barcodes and file sequences present in the data pool is essential for reliable data access. To overcome these limitations, previous approaches have used spatial segregation of data into distinct pools [16]. While PCR is typically known for its ease of amplifying specific DNA sequences, errors in priming via strand crosstalk can lead to information loss. In addition, selective amplification of a specific file using PCR requires access to the entire data pool for each query, which is also destructive to the sample queried. Finally, PCR-based approaches do not allow for physical deletion of specific files from a data pool, other than implementing an address overwrite [10].

As an alternative to PCR-based data access, inspired by genomic segmentation within biological cells, here we physically encapsulate and thereby isolate DNA-based molecular data within discrete silica capsules, which we subsequently label to enable random access of the data pool via hybridization and subsequent optical selection. Each unit of information encoded in DNA

we term a *file*, which includes both the DNA encoding the main data as well as any additional components used for addressing, storage, and retrieval. Each file contains a *file sequence*, consisting of the DNA encoding the main data, and *addressing barcodes*, or simply *barcodes*, which are additional short DNA sequences used to identify the file in solution using hybridization. We refer to a collection of files as a *data pool* or *database*, and the procedures for storing, retrieving, and reading out files is termed a *file system* (see **Supplementary Section S0** for a full list of terms).

As a proof-of-principle of our file system, we encapsulated 3,000-nt plasmids encoding 85-byte images, the files, within monodisperse, 6-μm spherical silica particles that were chemically surface-labeled using up to three 25-mer single-stranded DNA (ssDNA) oligonucleotides, the barcodes, chosen from a library of 240,000 orthogonal primers, allowing identification of up to ~$10^{15}$ possible distinct files using only three unique barcodes per file [17] (**Fig. 1**). Twenty icon-resolution images were chosen in the data pool to represent diverse subject matter including animals, plants, transportation, and buildings, and labeled with DNA barcodes that represent the categories to which each image belongs (**Supplementary Fig. 1**). Fluorescence-activated sorting (FAS) was used to select target subsets of the complete data pool by first annealing fluorescent oligonucleotide probes that are complementary to the barcodes, in order to address the DNA database [18]. Retrieval of specific, individual files and collections of files described by Boolean AND, OR, and NOT logic was achieved using combinations of distinct barcodes to query the data pool. Because physical encapsulation separates file sequences from barcodes used to describe the encapsulated information, our file system offers highly specific, robust data retrieval operations; the ability to delete specific subsets of data; in addition to long-term environmental protection of encoded file sequences via silica encapsulation [9,19,20]. While we apply our proposed file system to

a prototypical kilobyte-scale image database here, our approach is fully scalable to massive molecular data pools at the exabyte- and larger-scales, as well as alternative encapsulation strategies [21,22], barcode implementations [23-27], and physical or other sorting strategies using biochemical affinity, optical, or other labeling approaches [28-30].

**File Synthesis**

Digital information in the form of 20 icon-resolution images was stored in a data pool, with each image encoded into DNA and synthesized on a plasmid. We selected images of broad diversity, representative of distinct and shared subject categories, which included several domestic and wild cats and dogs, US presidents, and several human-made objects such as an airplane, boats, and buildings (**Fig. 1** and **Supplementary Fig. 1**). To implement this image database, the images were substituted with black-and-white, 26 × 26-pixel images to minimize synthesis costs, compressed using run-length encoding, and converted to DNA (**Supplementary Fig. 1**, **2**). Following synthesis, bacterial amplification, and sequencing validation (**Supplementary Fig. 3**), each plasmid DNA was separately encapsulated into silica particles containing a fluorescein dye core and a positively charged surface [19,20]. Because the negatively charged phosphate groups of the DNA interact with positively charged silica particles, plasmid DNA condensed on the silica surface, after which N-[3-(trimethoxysilyl)propyl]-N,N,N-trimethylammonium chloride (TMAPS) was co-condensed with tetraethoxysilane to form an encapsulation shell after four days of incubation at room-temperature [9,20] (**Fig. 2a**) to form discrete silica capsules containing the file sequence that encodes for the image file. Quantitative PCR (qPCR) of the reaction supernatant after encapsulation (**Supplementary Fig. 4**) showed full encapsulation of plasmids without residual DNA in solution. To investigate the fraction of capsules that contained plasmid DNA, we

compared the fluorescence intensity of the intercalating dye TO-PRO when added pre- versus post-encapsulation (**Supplementary Fig. 2**). All capsules synthesized in the presence of both DNA and TO-PRO showed a distinct fluorescence signal, consistent with the presence of plasmid DNA in the majority of capsules, compared with a silica particle negative control that contained no DNA. In order to test whether plasmid DNA was fully encapsulated versus partially exposed at the surface of capsules, capsules were also stained separately with TO-PRO post-encapsulation (**Fig. 2b**). Using qPCR, we estimated $10^6$ plasmids per capsule assuming quantitative recovery of DNA post-encapsulation (**Supplementary Fig. 5**).

Next, we chemically attached unique content addresses on the surfaces of silica capsules using orthogonal 25-mer ssDNA barcodes (**Supplementary Fig. 6**) describing selected features of the underlying image. For example, the image of an orange tabby house cat (**Supplementary Fig. 1**) was described with *cat*, *orange*, and *domestic*, whereas the image of a tiger was described with *cat*, *orange*, and *wild* (**Supplementary Fig. 1** and **Supplementary Table 2**). To attach the barcodes, we activated the surface of the silica capsules through a series of chemical steps. Condensation of γ-aminopropyltriethoxysilane with the hydroxy-terminated surface of the encapsulated plasmid DNA provided a primary amine chemical handle that supported further conjugation reactions (**Fig. 2c**). We modified the amino-modified surface of the silica capsules with β-azidoacetic acid N-hydroxysuccinimide (NHS) ester followed by an oligo(ethylene glycol) that contained two chemically orthogonal functional groups: the dibenzocyclooctyne functional group reacted with the surface-attached azide through strain-promoted azide-alkyne cycloaddition while the NHS ester functional group was available for subsequent conjugation with a primary amine. Each of the associated barcodes contained a 5'-amino modification that could react with the NHS-ester groups on the surface of the silica capsules, thereby producing the complete form

of our file. Notably, the sizes of bare, hydroxy-terminated silica particles representing capsules without barcodes were comparable with complete files consisting of capsules with barcodes attached, confirmed using scanning electron microscopy (**Fig. 2d** and **2e**, left). These results were anticipated given that the encapsulation thickness was only on the order of 10 nm [20] and that additional steps to attach functional groups minimally increases the capsule diameter. We also observed systematic shifts in the surface charge of the silica particles as different functional groups were introduced onto their surfaces (**Fig. 2e**). Using hybridization assays with fluorescently-labelled probes [31-33], we estimated the number of barcodes available for hybridization on our files to be on the order of $10^8$ (**Supplementary Fig. 7**). Following synthesis, files were pooled and stored together for subsequent retrieval. Illumina MiSeq was used to read each file sequence and reconstruct the encoded image following selection and de-encapsulation, in order to validate the complete process of image file encoding, encapsulation, barcoding, selection, de-encapsulation, sequencing, and image file reconstruction (**Supplementary Figs. 9**, **10**).

## File Selection

Following file synthesis and pooling, we used FAS to select specific targeted file subsets from the entire data pool. All files contained a fluorescent dye, fluorescein, in their core as a marker to distinguish files from other particulates such as spurious silica particles that nucleated in the absence of a core or insoluble salts that may have formed during the sorting process. Each detected fluorescein event was therefore interpreted to indicate the presence of an individual file at sufficiently low concentrations queried using FAS (**Supplementary Fig. 11**). For any query applied to the entire image database, a fluorescently-labelled ssDNA probe hybridized to its complementary barcode displayed externally on the surface of the silica capsule (**Fig. 3a**).

We subjected the entire data pool to a series of experiments to test selection sensitivity of target subsets using distinct queries. First, we evaluated single-barcode selection of an individual file, specifically *Airplane*, out of a pool of varying concentrations of the nineteen other files as background (**Fig. 3b**). To select the *Airplane* file, we hybridized an AFDye 647-labelled ssDNA probe that is complementary to the barcode *flying*, which is unique to *Airplane*. We were able to detect and select the desired *Airplane* file through FAS even at a relative abundance of $10^{-6}$ compared with each other file (**Fig. 3c**)**.** Comparison of the retrieved sequences between the flying gate and the NOT flying gate after chemical release of the file sequences from silica encapsulation revealed that 60–95% of the *Airplane* files were sorted into the flying gate (**Supplementary Figs. 18–21**). Note that any sort probability above 50% indicates enrichment of *Airplane* within the correct population subset (flying) relative to the incorrect subset (NOT flying), while a sort probability of 100% would indicate ideal performance.

**Boolean Search**

Aside from selecting single files, Boolean logic can be used to select a specific subset of the data pool. We demonstrated AND, OR, and NOT logical operations by first adding to the data pool fluorescently-labelled ssDNA probes that were complementary to the barcodes (**Fig. 4**, left). This hybridization reaction was used to distinguish one or several files in the data pool, which were then sorted using FAS. We used two to four fluorescence channels simultaneously to create the FAS gates that executed the target Boolean logic queries (**Fig. 4,** middle). To demonstrate a NOT query, we added to the data pool an AFDye 647-labelled ssDNA probe that hybridized to files that contained the *cat* barcode. Files that did not show AFDye 647 signal were sorted into the NOT cat subset (**Fig. 4a**). An example of an OR gate was applied to the data pool by simultaneously adding

*dog* and *building* probes that both had the TAMRA label (**Fig. 4b**). All files that showed TAMRA signal were sorted into the dog OR building subset by the FAS. Finally, an example of an AND gate was achieved by adding *fruit* and *yellow* probes that were labelled with AFDye 647 and TAMRA, respectively. Files showing signal for both AFDye 647 and TAMRA were sorted into the fruit AND yellow subset in the FAS (**Fig. 4c**). For each example query, we validated our sorting experiments by releasing the file sequence from silica encapsulation and sequencing the released DNA with Illumina MiniSeq (**Fig. 4**, right). Sort probabilities of each file for each search query are shown in **Supplementary Figs. S22–S24**.

The preceding demonstrations of Boolean logic gates enable sorting of files with varying specificity of selection criteria for the retrieval of different subsets of the data pool. FAS can also be used to create multiple gating conditions simultaneously, thereby increasing the specificity of file selections. To demonstrate increasingly complex Boolean search queries, we selected the file containing the image of Abraham Lincoln from the data pool, which included images of two presidents, George Washington and Abraham Lincoln. The *president* ssDNA probe, fluorescently-labeled with TAMRA, selected both *Lincoln* and *Washington* files from the data pool. The simultaneous addition of the *18$^{th}$ century* ssDNA probe, fluorescently-labeled with AFDye 647 (**Fig. 5a**, left), discriminated *Washington*, which contained the *18$^{th}$ century* barcode, from the *Lincoln* file (**Fig. 5a**, middle). The combination of these two ssDNA probes permitted the complex search query president AND (NOT 18$^{th}$ century). Sequencing analysis of the gated populations after reverse encapsulation validated that the sorted populations matched search queries for president AND (NOT 18$^{th}$ century), president AND 18$^{th}$ century, and NOT president (**Fig. 5a**, right; **Supplementary Fig. 25**).

To demonstrate the possibility of performing Boolean search using more than three fluorescence channels for sorting, we selected the *Wolf* file from the data pool using the query dog AND wild, and used the *black & white* probe to validate the selected file (**Fig. 5b**, left). Because conventional FAS software is only capable of sorting using 1D and 2D gates, we first selected one out of the three possible 2D plots (**Fig. 5b**, left and bottom): *dog*-TAMRA against *wild*-AFDye 647. We examined the *black & white*-TYE705 channel on members of the dog AND wild subset (**Fig. 5b**, left and bottom). Release of the encapsulated file sequence and subsequent sequencing of each gated population from the *dog* versus *wild* 2D plot validated sorting (**Fig. 5b**, right; **Supplementary Fig. 26**).

The use of plasmids as a substrate for encoding information offered a convenient workflow for restoring files into the data pool after retrieval. In cases where single images were sorted (**Figs. 4c**, **5a**, **b**), we were able to transform competent bacteria from each search query that resulted in a single file (**Supplementary Fig. 27**). Amplified material was pure and ready for re-encapsulation into silica particles, which could be re-introduced directly back into the data pool. Importantly, our molecular file system and file selection process thereby represents a complete write-access-read cycle that can in principle be applied to exabyte and larger-scale datasets. While sort probabilities were typically below the perfect 100% targeted for a specific file or file subset query, future work would be required to better characterize sources of error that may be due to sample contamination, FAS error, or imperfect orthogonality of barcode sequences employed (**Supplementary Fig. 6**) [17].

## Outlook

We present a non-destructive molecular file system that is capable of both specific file selection and Boolean logic search operations for random access of single files or file subsets in a data pool.

Our implementation easily scales by increasing the numbers of barcodes per file and query fluorophores used for file selection, which can thereby address files in a larger-scale database for random access and computation. For example, labeling each file using four distinct barcodes instead of only the three used here renders it possible to label $\binom{2.4\times10^5}{4} \approx 10^{20}$ files uniquely using the existing pool of ~$10^5$ orthogonal barcodes [17]. Assuming an FAS system is capable of sorting a single file from $10^6$ others using each fluorescent channel alone, as demonstrated in this work using a commercial FAS, one may theoretically sort a single file from $10^{24}$ others using a conventional four-channel FAS system. This file system would then in principle offer sufficient sensitivity and specificity to select a single file from an exabyte or even yottabyte data pool. However, the time needed to perform FAS scales linearly with the size of the data pool, which may be prohibitively long even for exabyte-scale data pools using only 10–100 bytes per file. For example, 12 minutes was required to select at least one hundred copies of the *Airplane* file in a data pool in which this file has a relative abundance of $10^{-6}$ compared with other files (**Fig. 3**). This is in contrast to selecting one hundred copies of the *Airplane* file in a data pool that contained equivalent numbers of nineteen other files, which required only ~30 seconds. Thus, in order to search through an entire exabyte-scale data pool within 24 hours, each file should consist of approximately 100 gigabytes, assuming a typical commercial FAS device that searches at 10,000 files per second. In order to reduce file selection time, future implementations of our molecular file system should therefore leverage parallel microfluidics-based optical sorting procedures and brighter fluorescence probes to increase selection throughput and sensitivity, and thereby reduce the pool search time. Alternatively, direct magnetic pulldown of files labelled with biochemical or affinity tags may be employed [30].

Aside from speed and specificity of data access, data density is also of importance to DNA data storage. Notably, both file size and data density can be tuned independently in our file system by changing the information content of loaded DNA and the size of the silica particles employed for encapsulation. While we used 6-µm silica core particles here in order to maximize fluorescence signal-to-noise ratios for a commercial FAS instrument, this also limited volumetric density of our DNA file system [3]. Specifically, using this approach an exabyte-scale data pool consisting of a 100-byte file per particle would require approximately $10^{16}$ files and 1 $m^3$ total dry volume, or $10^{18}$ bytes per $m^3$. In comparison, PCR-based random access has a theoretical volumetric density limit of $10^{24}$ bytes per $m^3$ [3], although additional methods are required to prevent crosstalk between file sequences and barcodes. To further increase the data density of our file system, future implementations may benefit from using nanoparticles ~100–200 nm in diameter to encode files [9,19,20] and higher sensitivity FAS systems [34,35] or direct biochemical, magnetic, or other pulldown for file and data subset selection from the data pool.

Beyond increasing file selection speed and data density, utilization of spectrally distinct fluorescent probes and discrete labeling intensities [26] would allow for far more complex and efficient logical operations than demonstrated here [36]. Physical particle parameters including forward and side-scatter could additionally be used to perform multi-dimensional sorting of particles with different scattering cross-sections, with or without additional fluorescence channels [37]. Repeated cycles of file selection in series could also further increase selection fidelity. While our technical approach differs significantly from approaches that rely on selective amplification for block selection [8,12,16], in which amplifications may reduce fidelity of file selection, PCR-based random access approaches will typically have faster read-write times because they forgo encapsulation and de-encapsulation steps required by our approach [9,19,20], which is therefore ideally

suited to long-term, archival data storage and retrieval with periodic file and barcode renewal. Aside from DNA data storage, population enrichment on our prototypical database of 20 unique files encoded in DNA plasmids with silica encapsulation and retrieval demonstrated using barcodes labels may alternatively be applied directly to biological DNA and other nanoscale sample management, such as genomic samples in biobanking or protein-encoding databases [38]. In either case, subsets of data or genomic sample pools may be enriched using Boolean AND, OR, and NOT logic, which complements existing PCR-based approaches. These operations enrich the capabilities of performing computation and sorting on underlying molecular data pools, moving us closer to realizing an economically viable, functional, massive molecular file and operating system [18,39,40].

## References

1 Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nature Reviews Genetics* **20**, 456-466, doi:10.1038/s41576-019-0125-3 (2019).

2 Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods* **11**, 499-507, doi:10.1038/nmeth.2918 (2014).

3 Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nature Materials* **15**, 366, doi:10.1038/nmat4594 (2016).

4 Palluk, S. *et al.* De novo DNA synthesis using polymerase-nucleotide conjugates. *Nature Biotechnology* **36**, 645-650, doi:10.1038/nbt.4173 (2018).

5 Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications* **10**, 2383, doi:10.1038/s41467-019-10258-1 (2019).

6 Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628, doi:10.1126/science.1226355 (2012).

7 Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77-80, doi:10.1038/nature11875 (2013).

8 Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A rewritable, random-access DNA-based storage system. *Scientific Reports* **5**, 14138, doi:10.1038/srep14138 (2015).

9 Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie International Edition* **54**, 2552-2555, doi:10.1002/anie.201411378 (2015).

10 Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Scientific Reports* **7**, 5011, doi:10.1038/s41598-017-05188-1 (2017).

11 Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950-954, doi:10.1126/science.aaj2038 (2017).

12 Organick, L. *et al.* Random access in large-scale DNA data storage. *Nature Biotechnology* **36**, 242–248, doi:10.1038/nbt.4079 (2018).

13 Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T. & Ohuchi, A. in *DNA 2002: DNA Computing.* (eds M. Hagiya & A. Ohuchi) 112–123 (Lecture Notes in Computer Science, Vol. 2568, Springer, 2003).

14 Yamamoto, M., Kashiwamura, S., Ohuchi, A. & Furukawa, M. Large-scale DNA memory based on the nested PCR. *Natural Computing* **7**, 335-346 (2008).

15 Yamamoto, M., Kashiwamura, S. & Ohuchi, A. in *DNA 2007: DNA Computing*. (eds M.H. Garzon & H. Yan) 99–108 (Lecture Notes in Computer Science, Vol. 4848, Springer).

16 Newman, S. *et al.* High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nature Communications* **10**, 1706 (2019).

17 Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proceedings of the National Academy of Sciences* **106**, 2289-2294, doi:10.1073/pnas.0812506106 (2009).

18 Reif, J. H. *et al.* in *DNA 2001: DNA Computing*. (eds N. Jonoska & N.C. Seeman) 231–247 (Lecture Notes in Computer Science, Vol. 2340, Springer, 2002).

19 Paunescu, D., Fuhrer, R. & Grass, R. N. Protection and deprotection of DNA--high-temperature stability of nucleic acid barcodes for polymer labeling. *Angewandte Chemie International Edition* **52**, 4269-4272, doi:10.1002/anie.201208135 (2013).

20 Paunescu, D., Puddu, M., Soellner, J. O. B., Stoessel, P. R. & Grass, R. N. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA "fossils". *Nature Protocols* **8**, 2440, doi:10.1038/nprot.2013.154 (2013).

21 Alexakis, T. *et al.* Microencapsulation of DNA within alginate microspheres and crosslinked chitosan membranes for in vivo application. *Applied Biochemistry and Biotechnology* **50**, 93-106 (1995).

22 Borodina, T. *et al.* Controlled release of DNA from self-degrading microcapsules. *Macromolecular Rapid Communications* **28**, 1894-1899, doi:10.1002/marc.200700409 (2007).

23 Braeckmans, K. *et al.* Encoding microcarriers by spatial selective photobleaching. *Nature Materials* **2**, 169-173, doi:10.1038/nmat828 (2003).

24 Wilson, R., Cossins, A. R. & Spiller, D. G. Encoded microcarriers for high-throughput multiplexed detection. *Angewandte Chemie International Edition* **45**, 6104-6117, doi:10.1002/anie.200600288 (2006).

25 Pregibon, D. C., Toner, M. & Doyle, P. S. Multifunctional encoded particles for high-throughput biomolecule analysis. *Science* **315**, 1393-1396, doi:10.1126/science.1134929 (2007).

26 Dagher, M., Kleinman, M., Ng, A. & Juncker, D. Ensemble multicolour FRET model enables barcoding at extreme FRET levels. *Nature Nanotechnology* **13**, 925-932, doi:10.1038/s41565-018-0205-0 (2018).

27 Martino, N. *et al.* Wavelength-encoded laser particles for massively multiplexed cell tagging. *Nature Photonics* **13**, 720-727, doi:10.1038/s41566-019-0489-0 (2019).

28 Lee, H., Kim, J., Kim, H., Kim, J. & Kwon, S. Colour-barcoded magnetic microparticles for multiplexed bioassays. *Nature Materials* **9**, 745-749, doi:10.1038/nmat2815 (2010).

29 Stewart, K. *et al.* in *International Conference on DNA Computing and Molecular Programming.* 55-70 (Vol. Springer).

30 Tomek, K. J. *et al.* Driving the scalability of DNA-based information storage systems. *ACS Synthetic Biology* **8**, 1241-1248, doi:10.1021/acssynbio.9b00100 (2019).

31 Pillai, P. P., Reisewitz, S., Schroeder, H. & Niemeyer, C. M. Quantum-dot-encoded silica nanospheres for nucleic acid hybridization. *Small* **6**, 2130-2134, doi:10.1002/smll.201000949 (2010).

32 Leidner, A. *et al.* Biopebbles: DNA-functionalized core–shell silica nanospheres for cellular uptake and cell guidance studies. *Advanced Functional Materials* **28**, 1707572, doi:10.1002/adfm.201707572 (2018).

33 Sun, P. *et al.* Biopebble containers: DNA-directed surface assembly of mesoporous silica nanoparticles for cell studies. *Small* **15**, 1900083, doi:10.1002/smll.201900083 (2019).

34 van Gaal, E. V. B., Spierenburg, G., Hennink, W. E., Crommelin, D. J. A. & Mastrobattista, E. Flow cytometry for rapid size determination and sorting of nucleic acid containing nanoparticles in biological fluids. *Journal of Controlled Release* **141**, 328-338, doi:10.1016/j.jconrel.2009.09.009 (2010).

35 Lian, H., He, S., Chen, C. & Yan, X. Flow cytometric analysis of nanoscale biological particles and organelles. *Annual Review of Analytical Chemistry* **12**, 389-409, doi:10.1146/annurev-anchem-061318-115042 (2019).

36 Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology* **4**, 648-655, doi:10.1038/nri1416 (2004).

37 Mage, P. L. *et al.* Shape-based separation of synthetic microparticles. *Nature Materials* **18**, 82-89, doi:10.1038/s41563-018-0244-9 (2019).

38 Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343-347, doi:10.1126/science.aao5167 (2018).

39 Baum, E. B. Building an associative memory vastly larger than the brain. *Science* **268**, 583-585 (1995).

40 Song, X. & Reif, J. Nucleic acid databases and molecular-scale computing. *ACS Nano* **13**, 6256-6268, doi:10.1021/acsnano.9b02562 (2019).

**Acknowledgments.** We gratefully acknowledge fruitful discussions with Charles Leiserson and Tao B. Schardl on the scalability and generalizability of our barcoding approach. We thank Glenn Paradis, Michael Jennings, and Michele Griffin of the Flow Cytometry Core at the Koch Institute in MIT and Patricia Rogers of the Flow Cytometry Facility at the Broad Institute of Harvard and MIT for assistance and fruitful discussions in developing the flow cytometry workflow. We also thank David Mankus of the Nanotechnology Materials Core Facility at the Koch Institute in MIT for assistance in the imaging of the particles using the scanning electron microscope and Alla Leshinsky of the Biopolymer and Proteomics Core at the Koch Institute at MIT for assistance in mass spectrometry.

**Funding.** M.B., J.L.B., T.R.S., and J.B. gratefully acknowledge funding from the Office of Naval Research N00014-17-1-2609, N00014-16-1-2506, N00014-12-1-0621, and N00014-18-1-2290 and the National Science Foundation CCF-1564025 and CBET-1729397. Additional funding to J.B. was provided through an NSF Graduate Research Fellowship (Grant # 1122374). P.C.B. was supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. C.M.A. was supported by NIH grant F32CA236425.

**Author contributions.** J.L.B., T.R.S., and M.B. designed the file labeling and selection scheme. J.L.B, T.R.S., and C.M.A. implemented the file selection scheme using FAS. J.B. and T.R.S. developed the encoding scheme and metadata tagging of the images to DNA. T.R.S. designed the plasmid for encoding imaging. H.H. and T.R.S. performed the cloning, transformation, and purification of the plasmids. J.L.B. synthesized and purified all the TAMRA and AFDye 647-

385 labelled DNA oligonucleotides. J.L.B. characterized the particles. J.L.B. developed the synthetic
386 route to attach DNA barcodes on the surface of the particles. J.L.B. performed the encapsulation,
387 barcoding, sorting, reverse encapsulation of the particles after sorting, and desalting. T.R.S., H.H.,
388 and M.R. performed the sequencing. J.B. performed computational validation of the orthogonality
389 of barcode sequences. J.B. developed the computational workflow to analyze the sequencing data,
390 including statistical analyses. M.B. conceived of the file system and supervised the entire project.
391 P.C.B. supervised the FAS selection and supervised the sequencing workflow. All authors
392 analyzed the data and equally contributed to the writing of the manuscript.

393 **Competing interests.** T.R.S., J.L.B., J.B. & M.B. have filed provisional patents (17/029,948 and
394 16/012,583) related to this work.

395 **Materials and correspondence.** Gene sequences and plasmid maps are available from AddGene
396 (https://www.addgene.org/depositing/77231/). Software for sequence encoding and decoding is
397 publicly available on GitHub (https://github.com/lcbb/DNA-Memory-Blocks/). All the data files
398 used to generate the plots in this manuscript are available from M.B. upon request.

399 **Online content.** Any methods, additional references, and supplementary information are available
400 at https://doi.org/10.10XX/XXXXX.
401

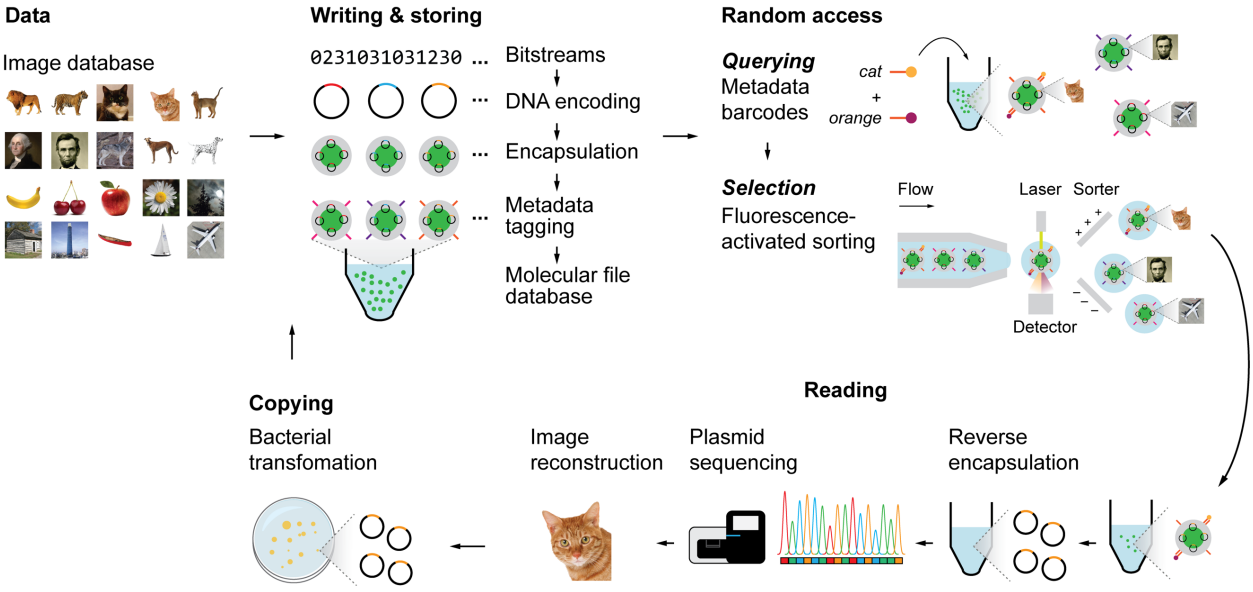
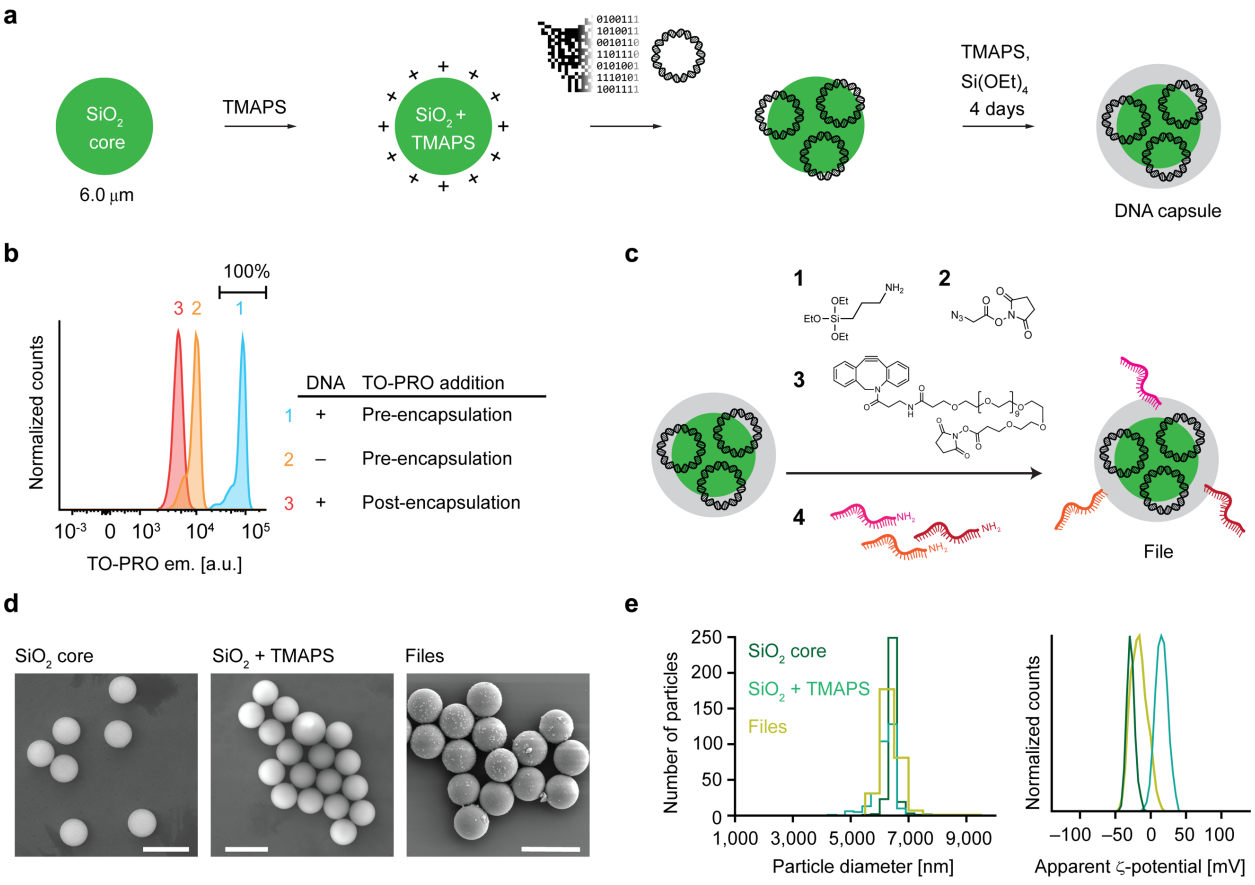



**Figure 1 | Write-access-read cycle for a content-addressable molecular filesystem.** Colored images were converted into 26 × 26-pixel, black-and-white icon bitmaps. The black-and-white images were then converted into DNA sequences using ternary encoding scheme [7]. The DNA sequences that encoded the images (file sequences) were inserted into a pUC19 plasmid vector and encapsulated into silica particles using sol-gel chemistry. Silica capsules were then addressed with content barcodes using orthogonal 25-mer single-stranded DNA strands, which were the final forms of the files. Files were pooled to form the molecular file database. To query a file or several files, fluorescently-labelled 15-mer ssDNA probes that are complementary to file barcodes were added to the data pool. Particles were then sorted with fluorescence-activated sorting (FAS) using two to four fluorescence channels simultaneously. Addition of a chemical etching reagent into the sorted populations released the encapsulated DNA plasmid. Sequences for the encoded images were validated using Sanger sequencing or Illumina MiniSeq. Because plasmids were used to encode information, re-transformation of the released plasmids into bacteria to replenish the molecular file database thereby closed the write-access-read cycle.




418

419

420 **Figure 2 | Encapsulation of DNA plasmids into silica and surface barcoding. a,** Workflow of
421 silica encapsulation [20]. **b**, Raw fluorescence data from FAS experiments to detect DNA staining of
422 TO-PRO during or after encapsulation. **c**, Functionalization of encapsulated DNA particles. **d,**
423 Scanning electron microscopy images of bare silica particles, silica particles functionalized with
424 TMAPS, and the file. **e,** Distribution of particle sizes determined from microscopy data (left) and
425 zeta potential analyses of silica particles and files.

426

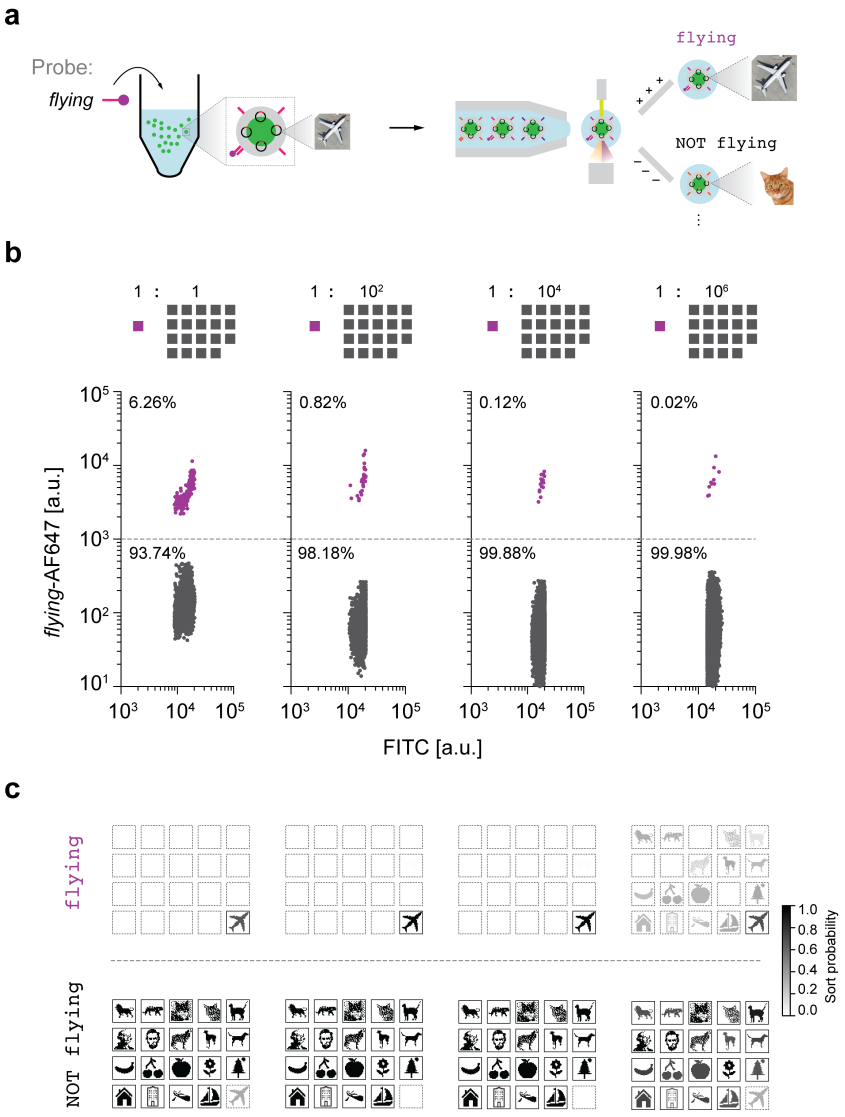



**Figure 3 | Single-barcode sorting. a,** Schematic diagram of file sorting using FAS. **b**, Sorting of *Airplane* from varying relative abundance of the other nineteen files as background. Percentages represent the numbers of particles that were sorted in the gate. Colored traces in each of the sorting plots indicate the target population. **c,** Sequencing validation using Illumina MiniSeq. Sort probability is the probability that a file is sorted into one gated population over the other gated populations. Boxes with solid outlines indicate files that should be sorted into the specified gate. Other files have dashed outlines.

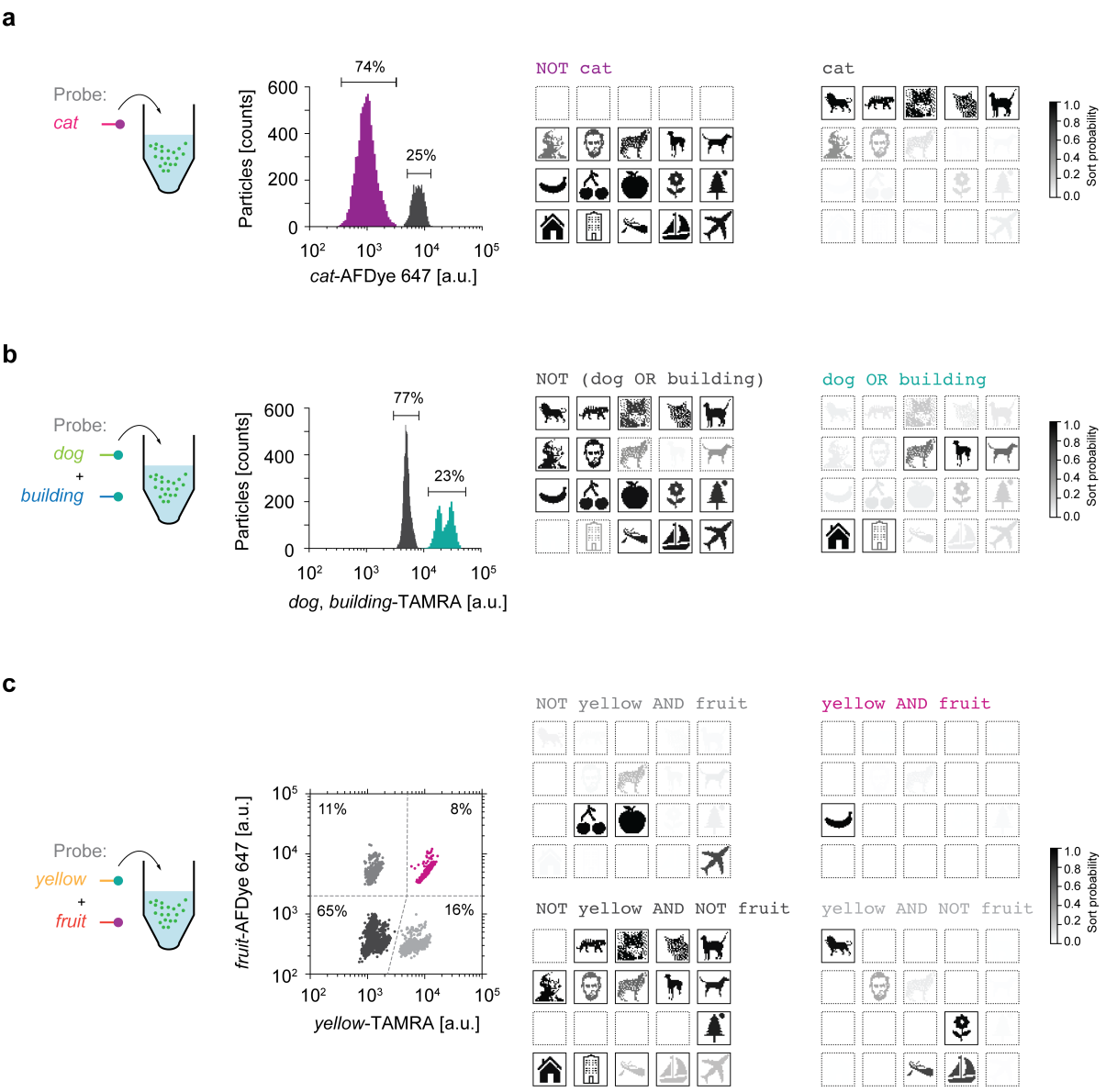



**Figure 4 | Fundamental Boolean logic gates. a,** `NOT cat` selection. Raw fluorescence trace from the FAS system (left) plotted on a 1D sorting plot showing the percent of particles that were sorted in each gate. Sequencing using Illumina MiniSeq tested selection specificity (right). **b,** `dog OR building` selection. Raw fluorescence trace from the FAS system (left) plotted on a 1D sorting plot showing the percent of particles that were sorted in each gate. Sequencing using Illumina MiniSeq evaluated sorting using the OR gate (right). **c,** A 2D sorting plot to perform a

444 `yellow AND fruit` gate. Percentages in each quadrant show the percentages of particles that
445 were sorted in each gate. Colored traces in all of the sorting plots indicate the target populations.
446 Sort probability is the probability that a file is sorted into one gated population versus the other
447 gated populations. Boxes with solid outlines indicate files that were intended to sort into the
448 specified gate. Other files have dashed outlines.
449

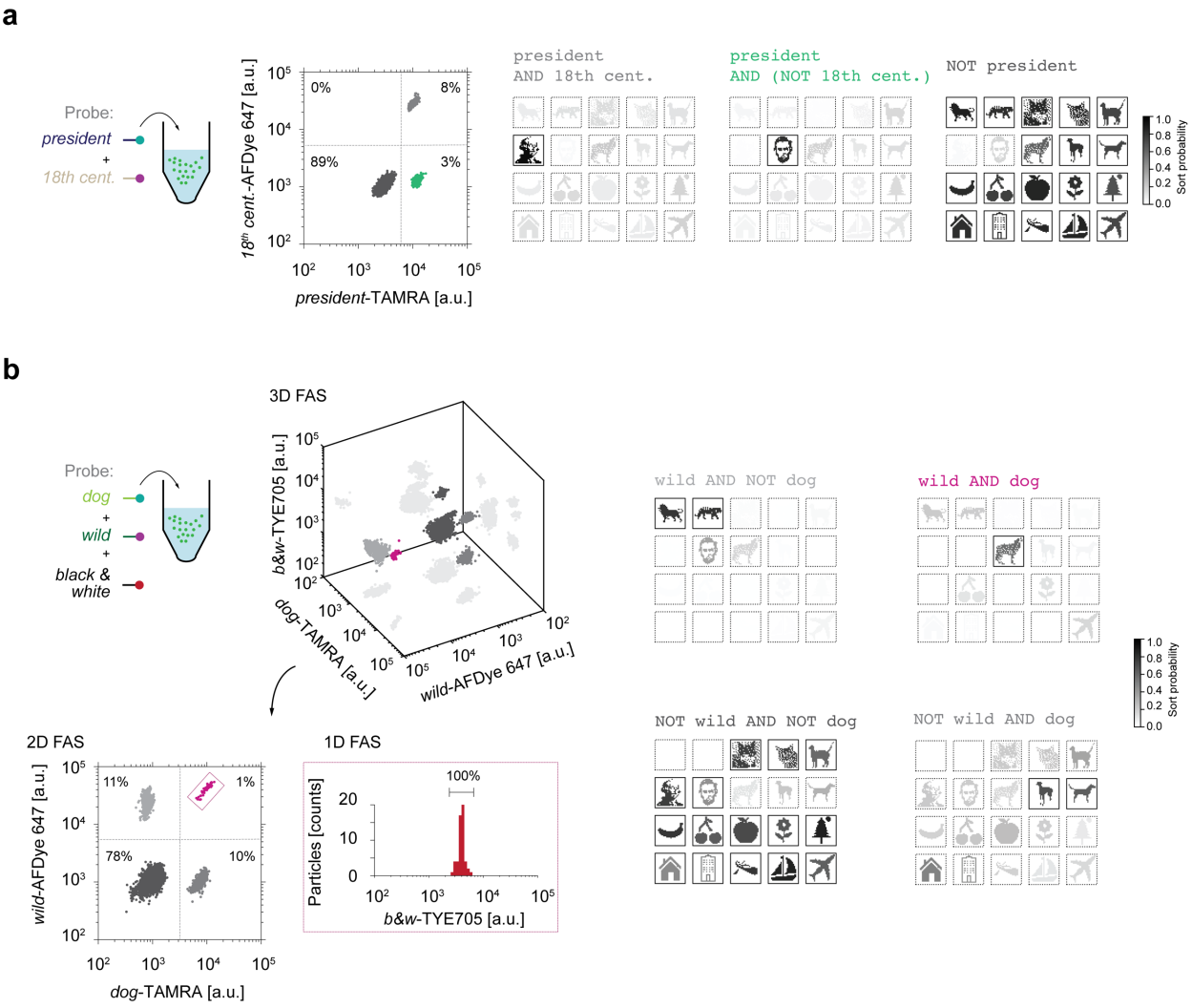



**Figure 5 | Arbitrary logic searching. a,** `president AND (NOT 18th century)` sorting. A 2D sorting plot (middle) was used to sort *Lincoln* by selecting a population that has high TAMRA fluorescence but low AFDye 647 fluorescence. Sequencing using MiniSeq offered quantitative evaluation of the sorted populations. **b,** Multiple fluorescence channels were projected into a 3D FAS plot (left and top). There are three possible 2D plots that can be used for sorting. To select the *Wolf* image using the query wild AND dog, a 2D plot of *wild* versus *dog* was first selected and then populations selected using quadrant gates (left and bottom). One of the quadrants were then selected where the *Wolf* image should belong based on the `wild AND dog` query in

460 order to test whether only a single population was present in the TYE705 fluorescence channel.
461 Sequencing quantified the sorted populations (right) using Illumina MiniSeq. Sort probability is
462 the probability that a file was sorted into one gated population over the other gated populations.
463 Boxes with solid outlines indicate files that would ideally be sorted into the specified gate. Other
464 files have dashed outlines.
465