High-throughput, Scalable, Quantitative, Cellular Phenotyping using X-Ray Tomographic Microscopy

Kevin Mader [1,2], Leah-Rae Donahue [3], Ralph Müller [4], and Marco Stampanoni [1,2]

1. Swiss Light Source, Paul Scherrer Institut, 5232 Villigen, Switzerland
2. Institute for Biomedical Engineering, University and ETH Zurich, 8092 Zurich, Switzerland
3. The Jackson Laboratory, Bar Harbor, ME, United States
4. Institute for Biomechanics, ETH Zurich, 8093 Zurich, Switzerland

kevin.mader@psi.ch
lrd@jax.org
ram@ethz.ch
marco.stampanoni@psi.ch

Abstract. With improvements in rate and quality of deep sequencing, the bottleneck for many genetic studies has become phenotyping. The complexity of many biological systems makes even developing these phenotypes a challenging task. In particular cortical bone can contain 10s of thousands of osteocyte cells interconnected in a complicated network. Easily measurable ensemble phenotypes like average size and density describe only a small portion of the variation in the system. We demonstrate a new approach to high-throughput phenotyping using Synchrotron-based X-ray Tomographic Microscopy (SRXTM) combined with our custom 3D image processing pipeline known as TIPL. The cluster-based evaluation tool enables high-speed data exploration and hypothesis testing over millions of structures. With these tools, we compare different strains of mice and look for trends in millions of cells. The flexible infrastructure offers a full spectrum of shape, distribution, and connectivity metrics for cellular networks and can be adapted to a wide variety of new studies requiring high sample counts such as the drug-gene interactions.

Keywords: phenotyping, high-throughput, screening, tomography, morphology, cellular networks, big data

1 Introduction

The networks formed by groups of cells play a significant role in nearly all biological systems ranging from small multicellular worms to the human nervous system. The function of these networks ranges from the more menial tasks of nutrient and waste transport to complicated signaling pathways. Functionally

they are essential for the development and function of larger cellular systems. In smaller systems the network may consist of several dozen cells, while in larger organisms there can be hundreds of millions of cells and even more possible connections between them. The scale of the measurements and consequential analysis is daunting and easily exceeds the capabilities of standard desktop tools. Furthermore with thousands of different cells in each specimen it is common that the variance within a sample for a given phenotype is larger than those between groups[1–6] making further analysis such as genetic trait localization difficult.

Social networks like Facebook, Google+, and LinkedIn have already encountered problems of this magnitude having in excess of 1.1 billion active monthly users (`http://newsroom.fb.com/content/default.aspx?NewsAreaId=22`) with an average of 190 connections for each user (`https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859`). They have consequently developed a series of tools which belong to the movement collectively known as "Big Data". This movement encompasses an entire class of problems where the standard desktop tools become overwhelmed because the volume, heterogeneity, or rate at which the data comes in is too high. The tools developed allow these companies to analyze, explore, and perform hypothesis testing on very large sets of data in order to capitalize from the information.

Any process can only occur as viable as its rate-limiting step. As a corollary, processes that are strongly limited by a single step are disproportionately improved by improvements in that single step. More broadly this means, the cumulative effect of a steady improvement in many different fields is, a rapid paradigm shift once the weakest link has been improved. In computing, this has been seen multiple times as a chip made a computer cheap enough to be in a home they appeared everywhere, again when the possibility to collect and track users on the internet went from a couple of hundred megabytes a year to petabytes. In genetics, this is being seen as the cost of sequencing a genome dropped from $3 billion in 2000 to $10,000 in 2010 [7]. The decreased cost of sequencing has moved the rate-limiting step further down the chain. In many areas, the task of accurately defining and measuring complicated phenotypes can be significantly more time-consuming than the sequencing itself [8–11].

Looking specifically at the example of genomics, the breakdown of researcher time and energy has shifted radically due to the rapid improvement in techniques with regard to speed and cost [7]. The division of research time between conducting experiments and analyzing data has changed entirely, and consequently the desired skill sets in new biologists wishing to enter the field have gone from experimental to analytical. The field is additionally a good choice for further examination because the transition has been handled well and they have started examining and developing solutions to the series of challenges such a transition brings on [12]. We believe, 3D tomographic imaging has finally reached this tipping point as well. Inside a 3-year period, the time to acquire a single scan has dropped by multiple orders of magnitude from many minutes to fractions of a second. Thus like the field of genetics, the division of researchers' time on many

projects has been shifted radically away from the standard break up (fig. 1). The time spent acquiring data is now minuscule compared to the time for data post-processing and analysis. The rate-limiting step now has shifted from the researcher's ability to conduct the experiment to the ability to analyze the data. This change is even more pronounced when looking at cellular networks where tens of thousands of cells can be measured in a single sample [6]. Furthermore in fields like light-field microscopy [13] and new wide angle cameras [14] can measure data at equally starting rates.

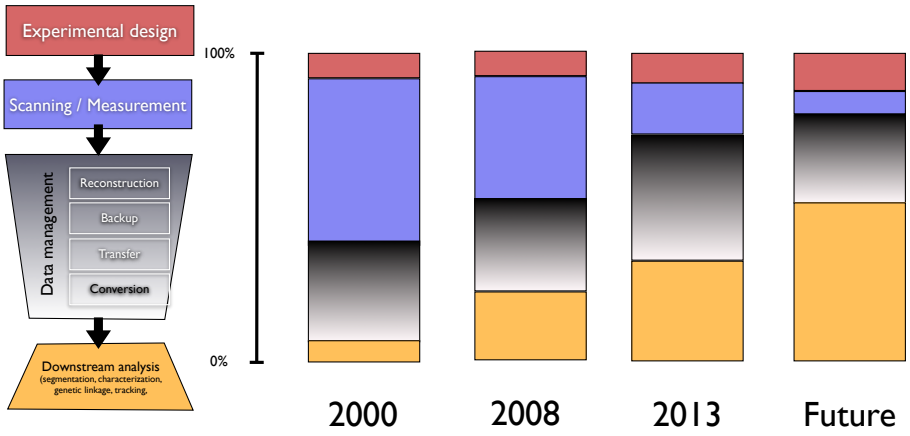

Fig. 1. Here is the researcher time breakdown analysis (inspired by [7]) applied to tomography and 3D imaging experiments. The colors on the bar graphs represent approximate proportion of researchers' time for each of the different aspects: Experimental design, measurement, data management, and downstream / post-processing analysis. The fourth column is how we expect the field to change over the coming years based on our experience with 100s of users.

The tools developed in this manuscript can begin to alleviate this issue and rebalance the division of time. While nothing is future-proof, an important question for every new toolset or framework is how will it handle the changes that come with time. To address this question we examined how companies and other groups have handled similar issues. A worldwide phenomenon known as "Big Data" [15–18, 12, 7] is a very loosely defined term, but generally refers to an increase in the volume (total size), velocity (data rate), and variety of data to be processed, which are beyond the capabilities of standard hardware and software approaches. Many software companies reached and far exceeded the projected usage, specifically services like YouTube on Google currently process 72 hours of new uploaded video every minute (http://www.reelseo.com/youtube-statistics-growth-2012/) or roughly 10-20 Gigabytes per second continuously. The primary tool used at Google for processing this volume of data is general framework called

MapReduce [19], which allows large complex jobs to be reliably distributed over thousands of computers. Other sites like Instagram (with 100 million users and 5 billion images) make use of cloud computing to automatically scale to the current demands of the site [20] .

In this manuscript, we show the tools and approach used to analyze cellular networks in bones and how our framework allows cellular networks to be analyzed with the same advanced tools used to examine social networks on a much larger scale [21]. The methodology can be applied to any number of different types of cellular networks measured with any 3D imaging modality. Furthermore the methods enable us to dig deeper into the data and explore it as a whole dataset rather than in the summarized views offered by standard database tools. Using tools like K-Means clustering [22] , Random Forests [23], Linear Discriminent Analysis [24], and Principal Component Analysis [25] new more exact phenotypes can be extracted from the data which more aptly describe the differences between groups.

Robust, flexible, and automated solutions substantially reduce the opportunity for unintentional user or system errors to slip into the results. With all the analyses run with the same tool and all of the results stored in a central database, the likelihood for user error stemming from the manual manipulation of data and parameters, which silently plague science, are significantly reduced.

2 Methods

The high-throughput phenotyping pipeline consists of an automated sample exchange and alignment setup [26] paired with an image processing [27, 6] framework to segment and analyze the images. Our tools, built on top of Spark[28, 29], uses a MapReduce-style approach but benefits from more dynamism and real-time querying that allows us to explore and analyze millions of samples in seconds.

2.1 Biological Measurements

To obtain the cell networks, a small region in the cortical bone of murine femora was measured in over 1300 samples. The samples come from the second generation of a genetic cross between two strains of mice with high (C3H/HeJ) and low (C57BL/6J) bone mass. The samples were measured at the TOMCAT Beamline of the Swiss Light Source at the Paul Scherrer Institut in Switzerland. Using a sample exchange system the samples were automatically aligned [26]. The regions of interest (mid-diaphysis) were identified and scanned following the procedures described in [6].

2.2 Performance Analysis

The performance analysis was run by executing each command 10 times and calculating the average time per calculation.

2.3 Analysis

The analysis was performed using the software and hardware infrastructure described in section 6.1. K-Means analysis was performed to identify within the data and was done by selecting reasonable, complementary metrics such as lacuna stretch and oblateness and dividing into 2 groups. New phenotypes were thus made by tracking the percentage of lacuna in each sample which were classified in the first group. Principal component analysis was performed to optimize the variance by creating a linear combination of the existing phenotypes. The summarized data were exported as CSV files and then plotted within R [30] using the ggplot library[31].

3 Results

In total 1276 different animals were measured from the population. Automatic segmentation and analysis failed on only 4 of the samples where the alignment had not been successful and a significant portion of the sample was outside of the field of view. The full shape analysis resulted in 57 different metrics being measured for every cell of 35million cells.

3.1 Within and Between Sample Variation

The within (intra-) and between (inter-) sample variation are shown in the following table (table 1) and graphs (fig. 2 and 3). The results show that the variation inside each sample is very high and for most of the shown metrics higher than between all the samples in the group.

Phenotype	Within	Between	Ratio (%)
Length	36.97	4.28	864.08
Width	27.92	4.73	589.89
Height	25.15	4.64	542.55
Volume	67.85	12.48	543.74
Nearest Canal Distance	70.35	333.40	21.10
Density (Lc.Te.V)	144.40	27.66	522.10
Nearest Neighbors (Lc.DN)	31.86	1.84	1736.11
Stretch (Lc.St)	13.98	2.36	592.46
Oblateness (Lc.Ob)	141.27	18.46	765.08

Table 1. The results in the table show the within and between sample variation for selected phenotypes in the first two columns and the ratio of the within and between sample numbers (all as percentages). For differentiating samples the lower the better and 100% for the third column would indicate the differences between samples are the same magnitude as the differences within a sample.

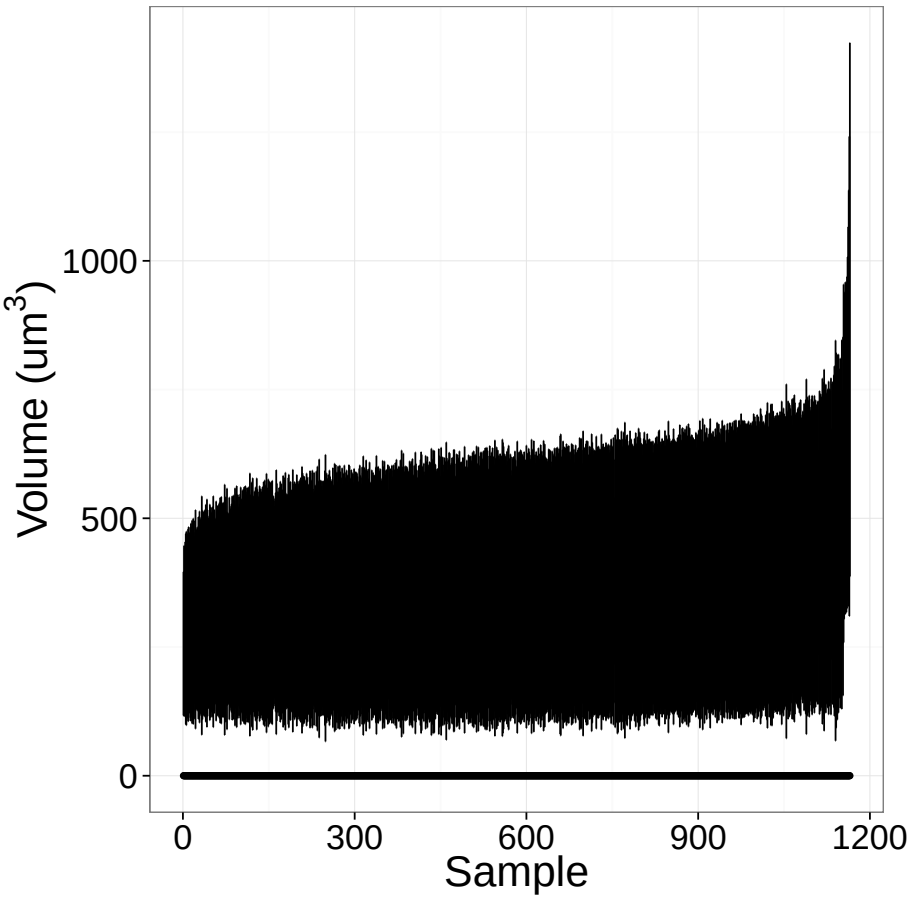

Fig. 2. The figure shows the volume in μm^3 (y-axis) against sample number(sorted by volume, x-axis). The point indicates the mean value inside the sample and the error bar shows the mean plus and minus the standard deviation.

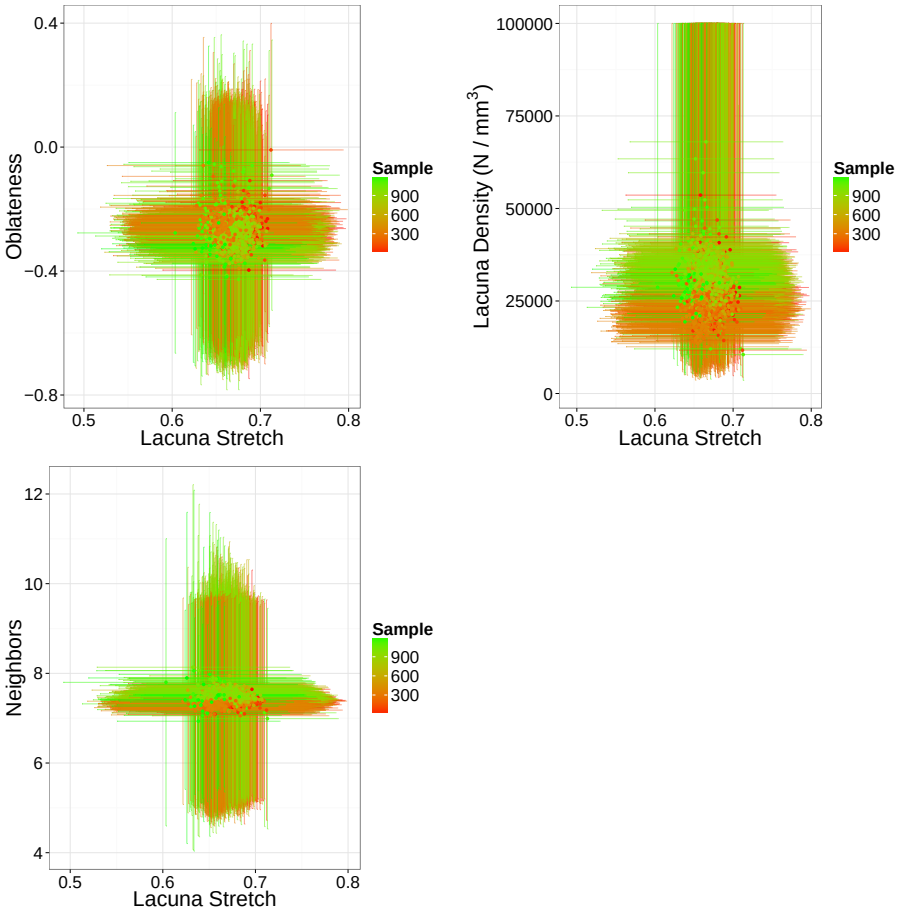

Fig. 3. The figure shows each samples as a point (colored by sample number) and the standard deviation of the given metrics as error bars in the x and y direction. The first panel shows the stretch against the oblateness. The second panel shows the stretch against the lacuna number density. The third panel shows the stretch against the neighbor count (Lc.DN).

3.2 New Phenotypes

The new phenotypes were developed using two different approaches: K-means and Principal Component analysis.

K-Means Clustering K-Means clustering was used to classify each lacunae into two groups based on stretch and oblateness. The resulting classification was saved and the number of lacuna in each group taken. The result for the analysis is summarized in figure 4, which shows how the new metric relates to existing phenotypes.

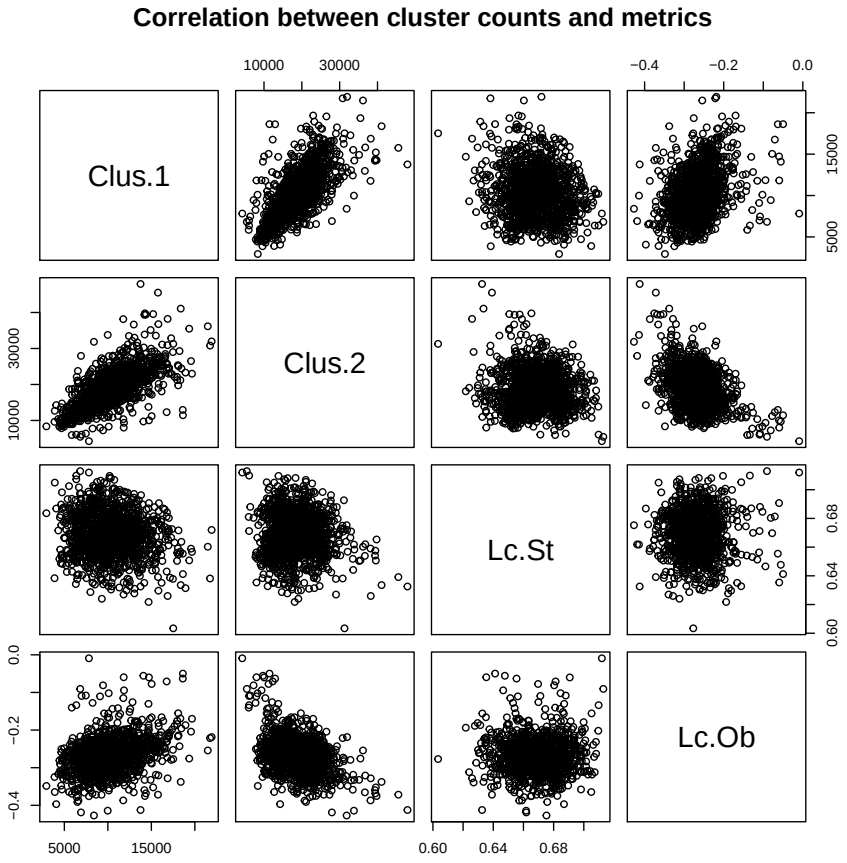

Fig. 4. The figure shows the new phenotypes based on K-means clustering plotted against the average values for the phenotypes used to generate it. Each column is a different metric on the X-axis and each row is a different metric on the y-axis. A high degree of correlation would be all points on the same non-flat line.

Shape A new phenotype was generated for shape using the lacuna stretch, volume, and oblatness metric into a principal component analysis (table. 2). The average and standard deviation were then calculated for each sample and summarized in the mean and variance table 3.

	PC1	PC2	PC3
Volume	0.45	0.82	0.36
Stretch	0.68	-0.05	-0.73
Oblateness	-0.58	0.57	-0.58

Table 2. The composition of the principal components, each column represents a component ordered by the largest to least contribution to total variance

Phenotype	Within	Between	Ratio (With./Bet.)
PrinComp 1	851.16	126.84	671.08
PrinComp 3	692.92	145.06	477.68

Table 3. The results in the table show the within and between sample variation for selected phenotypes in the first two columns and the ratio of the within and between sample numbers (all as percentages). 100% for the third column would indicate the differences between samples are the same magnitude as the differences within a sample and can be considered as the limit for clearly distinguishing samples from one another

Neighors / Density / Orientation A second principal component analysis was run on the Neighbor Count (Lc.ND), Density (Territory = Lc.Te.V), and orientation (vertical projection of principal orientation) information. The resulting components were well distributed between the 3 different metrics showing that each is contributing to the new phenotype. The output table (table 5) shows the

	PC1	PC2	PC3
Neighbor (Lc.DN)	-0.72	-0.10	0.69
Density (Lc.Te.V)	-0.68	0.33	-0.66
Orientation (Vertical)	-0.16	-0.94	-0.31

Table 4. The composition of the principal components, each column represents a component ordered by the largest to least contribution to total variance

ratios for each of these metrics.

Phenotype	Within	Between	Ratio (With./Bet.)
PrinComp 1	1520.52	131.09	1159.89
PrinComp 2	702.00	144.80	484.81
PrinComp 3	822.75	125.60	655.04

Table 5. The results in the table show the within and between sample variation for selected phenotypes in the first two columns and the ratio of the within and between sample numbers (all as percentages). 100% for the third column would indicate the differences between samples are the same magnitude as the differences within a sample and can be considered as the limit for clearly distinguishing samples from one another

3.3 Performance

The analysis was run using 40 cores spread between 2 standard nodes and 1 high-performance node on the cluster. To load and preprocess the results from 1276 comma separated text files took 10 minutes. Once the files were loading simple computations like computing the average of a given metric in the entire set took less than 400ms. A K-means analysis with 4 variables took less than 1s per iteration. By comparison loading the data on a single High Performance Node (sec. 6.1) took more than 6 hours and used 60GB of memory locally; a single column average took 4.6s and calculating an average volume grouped by sample took on average 47.8s. On machines with less memory, these operations would take significantly longer.

4 Conclusion

We have thus shown in this paper an approach for measuring and dealing with cellular network samples in a high-throughput manner. The tools enable us to manipulate and analyze data in an exploratory, scalable way without necessitating proprietary software, or particularly high performance supercomputers. Analyses such as the ones done in this paper can be done for well less than $100 using Amazon's EC2 cloud and can scale to many more thousands of samples as measurement techniques get faster and more detailed.

4.1 New Phenotypes

The K-means clustering provided new information non-correlated with other phenotypes when compared on the ensemble results. As it is a different type of metric it cannot be compared to the standard phenotypes in terms of with to between ratios, but it simplifies the data by reducing the number of different metrics to examine.

Using principal component analysis on the entire dataset enabled us to identify composite metrics which reduced the intra-to-inter sample variation below any of the composite parts (484% and 477% vs 543% and 522% respectively).

The reduction in this variation is substantial and makes further tasks like quantitative trait localization much easier since it focuses on the differences between samples. Furthermore looking at composition of these new phenotypes we can postulate at the potential underlying mechanisms which might cause them to vary less within a single sample.

4.2 Performance

While the processing is still possible using standard tools, such long delays for simple queries mean that it is more difficult to interactively explore the data and test hypotheses. Furthermore the costs of purchasing single computers capable of handling such datasets can be prohibitively expensive since both the processor count and memory demands are high. Distributed solutions based Java, Hadoop, and Spark can be very easily run on a large number of standard computers and automatically setup on many cloud-hosting services like Amazon making the barrier to entry very low. Furthermore due to the fault-tolerant design computers can crash or for Spark added during computations without interruption.

5 Outlook

The development of many of these tools is still in an early stage and while they automatically support a wide range of Java Libraries, the number of interactive statistical analysis, machine learning, and visualization options are limited when compared to more thoroughly developed platforms like Matlab (The Mathworks, Natick, MA) or R (R Foundation). There are many significant efforts being undertaken across the globe to further develop and increase accessibility for these tools and many of the existing shortcomings will likely be soon overcome. We showed in this paper using basic tools like Principal Component Analysis and K-Means clustering, but many other techniques are available for examining these datasets and the potential for discovering new underlying correlations and relationships is nearly limitless in such rich datasets. The ultimate goal of these techniques is to improve the number and quality of genes identified using techniques such as Quantitative Trait Localization. In the supplemental material we show some of the phenotypes compared with specific markers. The QTL analysis is an important next step for transforming these metrics into a better understanding of the underyling biological mechanisms.

References

1. Yasmin Carter, C David L Thomas, John G Clement, and David M L Cooper. Femoral osteocyte lacunar density, volume and morphology in women across the lifespan. *Journal of structural biology*, null(null), July 2013.
2. Edwin A Cadena and Mary H Schweitzer. Variation in osteocytes morphology vs bone type in turtle shell and their exceptional preservation from the Jurassic to the present. *Bone*, 51(3):614–20, September 2012.

3. Satoshi Hirose, Minqi Li, and Taku Kojima. A histological assessment on the distribution of the osteocytic lacunar canalicular system using silver staining. *Journal of bone and mineral metabolism*, 25(6):374–382, 2007.
4. Philipp Schneider, Martin Stauber, Romain Voide, Marco Stampanoni, Leah Rae Donahue, and R Müller. Ultrastructural Properties in Cortical Bone Vary Greatly in Two Inbred Strains of Mice as Assessed by Synchrotron Light Based Micro- and Nano-CT. *Journal for Bone Mineral Research*, 22(10):1557–1570, 2007.
5. Max Langer, Alexandra Pacureanu, Heikki Suhonen, Quentin Grimal, Peter Cloetens, and Françoise Peyrin. X-ray phase nanotomography resolves the 3D human bone ultrastructure. *PloS one*, 7(8):e35691, January 2012.
6. Kevin Scott Mader, Philipp Schneider, Ralph Müller, and Marco Stampanoni. A quantitative framework for the 3D characterization of the osteocyte lacunar system. *Bone*, 57(1):142–154, July 2013.
7. Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125, January 2011.
8. Natalie de Souza. High-throughput phenotyping. *Nature Methods*, 7(1):36–36, January 2010.
9. Christopher N Topp, Anjali S Iyer-Pascuzzi, Jill T Anderson, Cheng-Ruei Lee, Paul R Zurek, Olga Symonova, Ying Zheng, Alexander Bucksch, Yuriy Mileyko, Taras Galkovskyi, Brad T Moore, John Harer, Herbert Edelsbrunner, Thomas Mitchell-Olds, Joshua S Weitz, and Philip N Benfey. 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 110(18):E1695–704, April 2013.
10. D Ruffoni, T Kohler, R Voide, A J Wirth, L R Donahue, R Müller, and G H van Lenthe. High-throughput quantification of the mechanical competence of murine femora - A highly automated approach for large-scale genetic studies. *Bone*, 55(1):216–21, July 2013.
11. Jeffrey Jestes, Ke Yi, and Feifei Li. Building Wavelet Histograms on Large Data in MapReduce. pages 109–120, October 2011.
12. Lincoln D Stein. The case for cloud computing in genome informatics. *Genome biology*, 11(5):207, January 2010.
13. Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413–20, May 2013.
14. D J Brady, M E Gehm, R A Stack, D L Marks, D S Kittle, D R Golish, E M Vera, and S D Feller. Multiscale gigapixel photography. *Nature*, 486(7403):386–9, June 2012.
15. Shufen Zhang, Hongcan Yan, and Xuebin Chen. Research on Key Technologies of Cloud Computing. *Physics Procedia*, 33(null):1791–1797, January 2012.
16. Dinkar Sitaram and Geetha Manjunath. *Moving To The Cloud*, volume null. Elsevier, 2012.
17. Afsaneh Mohammadzaheri, Hossein Sadeghi, Sayyed Keivan Hosseini, and Mahdi Navazandeh. DISRAY: A distributed ray tracing by map-reduce. *Computers & Geosciences*, 52:453–458, March 2013.
18. Lizhe Wang, Jie Tao, Rajiv Ranjan, Holger Marten, Achim Streit, Jingying Chen, and Dan Chen. G-Hadoop: MapReduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems*, 29(3):750–739, October 2012.

19. Jeffrey Dean and Sanjay Ghemawat. MapReduce. *Communications of the ACM*, 51(1):107, January 2008.
20. SSDs Boost Instagram's Speed on Amazon EC2 - CIO.com.
21. Grzegorz Malewicz, Matthew H. Austern, Aart J.C Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel. In *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*, page 135, New York, New York, USA, June 2010. ACM Press.
22. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California, 1967.
23. Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.
24. C. Radhakrishna, Rao. The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B*, 10(2):159–203, 1948.
25. I.T. Jolliffe. *Principal Component Analysis (Springer Series in Statistics)*. Springer, 2002.
26. Kevin Mader, Federica Marone, Gordan Mikuljan, Andreas Isenegger, and Marco Stampanoni. High-throughput, fully-automatic, synchrotron-based microscopy station at TOMCAT. *Journal of Synchrotron Radiation*, 18(2):117–124, 2011.
27. Kevin Mader, Rajmund Mokso, and Christophe Raufaste. Quantitative 3D Characterization of Cellular Materials: Segmentation and Morphology of Foam. *Colloids and Surfaces A: ...*, 415(5):230–238, September 2012.
28. Michael J. Franklin Scott Shenker Ion Stoica Matei Zaharia, Mosharaf Chowdhury. Spark: Cluster computing with working sets.
29. Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. page 2, April 2012.
30. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
31. Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.
32. Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–81, July 2013.

6 Supplementary Material

6.1 Software Tools Used

1. Java (TM) SE Runtime Environment (build 1.7.0-25-b15)
2. Java HotSpot(TM) 64-bit Server VM (build 23.25-b01)
3. Spark 0.9.1
   (`https://github.com/apache/incubator-spark`
   Commit: 740e865f40704dc9158a6cf635990580fb6adcac)
4. Sun Grid Engine 6.2e5

6.2 Hardware Tools

1. Cluster Machines (Merlin4)

2. Standard Node
   one blade enclosure with 16 Xeon 5650/5670 12-core processors, total of 192 cores, 4 GB RAM/core
3. Fat Nodes
   one blade enclosure with 16 Xeon 5650/5670 12-core processors, total of 192 cores, 8 GB RAM/core (fat nodes)
4. High Performance Node
   one blade enclosure with Xeon E5-2670 16-core processor, 8 GB RAM/core
5. Network Interconnect
   4x QDR Infiniband for the compute nodes and the main cluster storage
6. Storage
   Main storage: GPFS on DDN S2A9900 hardware, 10GB of local storage.

Cluster Configuration The configuration used for the scripts on our cluster, while somewhat specific is publicly available at `https://github.com/skicavs/sge_spark`.

K-Means Calculations The K-Means calculations were done using a modified version of the built-in Spark script written in Scala (`https://github.com/apache/incubator-spark/blob/master/examples/src/main/scala/org/apache/spark/examples/SparkKMeans.scala`). The various versions of it can be requested from the authors.

Principal Component Analysis The principal component analysis was calculated using the scripts available from Thunder (`https://github.com/freeman-lab/thunder`) [32].

6.3 Phenotypes vs Genotypes

Here we compare the new phenotypes to several genotypes assessed using polymerase chain reaction (PCR) markers located on two different chromosomes.

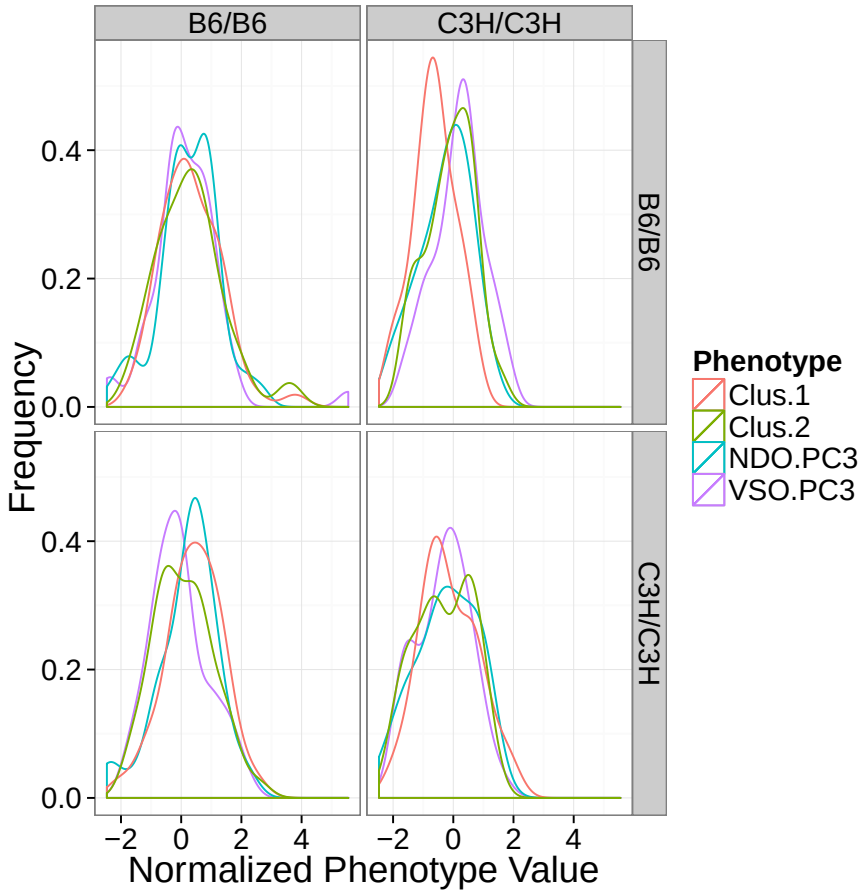

Fig. 5. The figure shows the new phenotypes based on K-means clustering and the principal component analysis. The values are plotted for using 2 different markers for genotype (D5Mit95 for the rows, and D9Mit259) for columns. Larger differences in distributions inducate the potential for a gene at this marker to be involved in the phenotype.

Table 6: The group comparison based on the genotype marker D5Mit95 located on the 5 chromosome.

	B6/B6 N=195	B6/C3H N=403	C3H/C3H N=182	p.overall	p.trend
Lc.V	363 (35.8)	366 (46.4)	391 (34.0)	<0.001	<0.001
Lc.St	0.67 (0.01)	0.67 (0.01)	0.68 (0.01)	<0.001	<0.001
Lc.Ob	-0.28 (0.04)	-0.28 (0.04)	-0.26 (0.04)	<0.001	<0.001
VSOb.PC1	-0.01 (0.15)	-0.01 (0.17)	0.06 (0.16)	<0.001	<0.001
VSOb.PC2	-0.01 (0.13)	-0.02 (0.16)	-0.05 (0.12)	0.020	0.021
NDO.PC1	0.00 (0.07)	0.00 (0.08)	0.02 (0.06)	0.008	0.027
NDO.PC3	0.00 (0.11)	0.00 (0.12)	0.02 (0.10)	0.065	0.234
Clus.1	10620 (2731)	10406 (2927)	11665 (2680)	<0.001	0.001
Clus.2	20460 (5151)	19832 (5473)	20153 (4687)	0.373	0.547

Table 7: The group comparison based on the genotype marker D9Mit259 located on the 9 chromosome.

	B6/B6 N=185	B6/C3H N=419	C3H/C3H N=191	p.overall	p.trend
Lc.V	382 (50.0)	371 (46.4)	358 (29.7)	<0.001	<0.001
Lc.St	0.67 (0.02)	0.67 (0.02)	0.67 (0.01)	0.257	0.163
Lc.Ob	-0.26 (0.05)	-0.28 (0.04)	-0.29 (0.05)	<0.001	<0.001
VSOb.PC1	0.00 (0.17)	0.01 (0.17)	0.01 (0.15)	0.895	0.647
VSOb.PC2	-0.02 (0.18)	-0.01 (0.15)	-0.03 (0.13)	0.248	0.265
NDO.PC1	0.02 (0.07)	0.01 (0.07)	-0.01 (0.08)	<0.001	<0.001
NDO.PC3	0.03 (0.11)	0.01 (0.11)	-0.03 (0.12)	<0.001	<0.001
Clus.1	11984 (2869)	10725 (2717)	9510 (2667)	<0.001	<0.001
Clus.2	21168 (5625)	20259 (5200)	18977 (4838)	<0.001	<0.001