Genetic Diversity in Chimpanzee Transcriptomics Does Not Represent Wild Populations

Navya Shukla[1,2], Bobbie Shaban[2,†], and Irene Gallego Romero[1,2,3,4,*]

[1]School of BioSciences, The University of Melbourne, Parkville, Victoria, Australia

[2]Melbourne Integrative Genomics, University of Melbourne, Parkville, Victoria, Australia

[3]Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Parkville, Victoria, Australia

[4]Center for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, Estonia

[†]Present address: Melbourne Data Analytics Platform (MDAP), The University of Melbourne, Parkville, VIC 3010, Australia

*Corresponding author: E-mail: irene.gallego@unimelb.edu.au.

Accepted: 1 November 2021

Abstract

Chimpanzees (*Pan troglodytes*) are a genetically diverse species, consisting of four highly distinct subspecies. As humans' closest living relative, they have been a key model organism in the study of human evolution, and comparisons of human and chimpanzee transcriptomes have been widely used to characterize differences in gene expression levels that could underlie the phenotypic differences between the two species. However, the subspecies from which these transcriptomic data sets have been derived is not recorded in metadata available in the public NCBI Sequence Read Archive (SRA). Furthermore, labeling of RNA sequencing (RNA-seq) samples is for the most part inconsistent across studies, and the true number of individuals from whom transcriptomic data are available is difficult to ascertain. Thus, we have evaluated genetic diversity at the subspecies and individual level in 486 public RNA-seq samples available in the SRA, spanning the vast majority of public chimpanzee transcriptomic data. Using multiple population genetics approaches, we find that nearly all samples (96.6%) have some degree of Western chimpanzee ancestry. At the individual donor level, we identify multiple samples that have been repeatedly analyzed across different studies and identify a total of 135 genetically distinct individuals within our data, a number that falls to 89 when we exclude likely first- and second-degree relatives. Altogether, our results show that current transcriptomic data from chimpanzees are capturing low levels of genetic diversity relative to what exists in wild chimpanzee populations. These findings provide important context to current comparative transcriptomics research involving chimpanzees.

Key words: chimpanzee, genetic diversity, RNA-seq, genotyping.

Significance

Chimpanzees are genetically much more diverse than humans. Currently categorized into four subspecies with high F_{ST} between them, they have much larger effective population size than humans and a complex past involving multiple admixture events. Using publicly available data from 486 RNA sequencing (RNA-seq) samples—the vast majority of all bulk RNA-seq chimpanzee data available on SRA—we characterize genetic diversity in current chimpanzee transcriptomic samples and find that transcriptomics falls very short of capturing extant genetic diversity in chimpanzees, with only 89 unique individuals in our sample, the majority of them from the same subspecies. We propose ways to ameliorate this unbalance, but given the status of chimpanzees as critically endangered, this will require proactive collaboration with zoos globally.

Introduction

Chimpanzees (*Pan troglodytes*) are one of two extant members of the genus *Pan*, estimated to have diverged from the ancestral human lineage between 6 and 13 Ma (Prado-Martinez et al. 2013). With a rich evolutionary history and high amounts of genetic diversity, they have long been used as experimental models in biological research, both as model systems and to further our understanding of human-unique traits (Bustamante et al. 2005; Varki and Altheide 2005; Coop and Przeworski 2007). After the sequencing of the chimpanzee genome showed that ~99% of sequence is common between chimpanzees and humans (ignoring genomic rearrangements), efforts have focused on identifying human-specific variation of potential functional significance (Waterson et al. 2005). Over the last 13 years, RNA sequencing (RNA-seq), a relatively low-cost and high-throughput technology, has emerged as one of the leading approaches in this question, with much work having been done to characterize differences in gene expression patterns across humans, chimpanzees, and other primates (Gallego Romero et al. 2012). Transcriptomic studies have compared gene expression across organ systems and provided chimpanzee transcriptomes for various tissues (Brawand et al. 2011; Cardoso-Moreira et al. 2019; Blake et al. 2020), often with a focus on the brain and neural development, to identify the molecular basis of cognitive and behavioral differences (Mostajo-Radji et al. 2020). In parallel, induced pluripotent stem cells (iPSCs) from chimpanzees, which can be derived from existing cell lines or donor animals through minimally invasive means, are rapidly becoming established as models to study developmental intermediates and other hard-to-sample tissues (Dannemann and Gallego Romero 2021).

The majority of this data is publicly accessible through the NCBI's Sequence Read Archive (SRA) (Kodama et al. 2011). However, not one of the 4,389 RNA-seq samples classified as *Pan troglodytes* in the SRA specifies the subspecies of donor individuals, suggesting that transcriptomic studies are not generally considering sample subspecies in their analyses. Yet chimpanzees show population stratification greater than all other lineages of great apes (Fischer et al. 2006; Prado-Martinez et al. 2013; de Manuel et al. 2016). They can be split into two monophyletic clades, each containing two distinct subspecies (Won and Hey 2004; Bjork et al. 2011; Prado-Martinez et al. 2013). The first clade consists of the closely related Central (*Pan troglodytes troglodytes*) and Eastern chimpanzees (*Pan troglodytes schweinfurthii*), found in equatorial Africa, and the second clade is comprised of the Western chimpanzees (*Pan troglodytes verus*), found in upper Guinea and the recently designated Nigeria-Cameroon (*Pan troglodytes ellioti*) chimpanzees from the Gulf of Guinea (Gonder and Disotell 2006; Oates et al. 2009; Prado-Martinez et al. 2013). Each subspecies has had a long separate history and has been exposed to a wide range of ecological variation, such as incidence of pathogens, imposing different selective pressure (Schmidt et al. 2019), although evidence also points to considerable gene flow between subpopulations (de Manuel et al. 2016; Lester et al. 2021). Pairwise F_{ST} between Central and Eastern chimpanzees is estimated to be 0.09, similar to values seen between different populations of humans, but F_{ST} between Western and Central chimpanzees is 0.29 and that between Western and Eastern is 0.32—significantly higher values than typical human estimates (Fischer et al. 2006). There is also extensive diversity within individual subspecies. For example, looking at a non-coding locus on the X chromosome shared between humans and chimpanzees, the mean pairwise sequence difference among all humans is 0.037%, whereas among Central chimpanzees alone it is 0.18% (Kaessmann et al. 1999).

In transcriptomic studies of chimpanzees, tissue samples are often collected post-mortem from captive individuals, following death from unrelated causes. However, it is unclear how representative of the diversity in wild chimpanzees these small captive populations are. A genetic survey looking at European zoos and research institutes showed that the majority of individuals (70%) had some Western ancestry (Hvilsom et al. 2013). A similar survey of American chimpanzees also produced highly skewed results, with 95% of studied individuals being of Western descent (Ely et al. 2005). Although collection strategies in studies often aim to be random, sampling from these reservoirs with depleted genetic diversity suggests that there is a strong likelihood that transcriptomic studies are not sampling much of the intra- and inter-subspecies diversity found in chimpanzees. Here, we consider that possibility, generating genotype data from publicly available chimpanzee RNA-seq samples to investigate the state of genetic diversity in current transcriptomic research involving chimpanzees through a population genetic lens.

Results

Initial Sample Selection and Genotyping

As of the 19th of January 2021, the NCBI SRA database contained 4,389 RNA-seq samples with taxa label *Pan troglodytes*. We focused on 3,221 samples sequenced by Illumina HiSeq 2000, 2500, and 4000 to reduce biases in the data caused by different sequencing technologies and instruments. After removing single-cell RNA-seq samples and further stringent filtering steps (see Materials and Methods), we retained 486 BioSamples from 40 different BioProjects (collections of BioSamples from a single initiative in NCBI; supplementary tables 1 and 2, Supplementary Material online) and 20 different tissue or cell types. These included 339 single-ended and 147 pair-ended samples; for fairness, we considered only R1 reads for paired-ended samples.

We note that in the SRA BioSample is an ambiguous term that covers different experimental contexts and designs.

Although in one study different BioSamples can refer to genetically distinct individuals, in another they may indicate different tissue sections or different experimental treatments applied to the same individual. For example, PRJNA299472 (S7, He et al. 2017) has 72 BioSamples—these correspond to 18 different sections from the prefrontal cortex of only 4 different individuals. In the following, we therefore distinguish *replicate* samples, which are BioSamples from the same donor animal and the same study, and *duplicate* samples, which are BioSamples from the same donor animal across different studies, defined on the basis of publicly available metadata either in SRA or the associated publication.

We successfully genotyped 6,943,957 SNPs in these 486 samples, 54,706 of which had ≤5% missingness. An initial UPGMA tree (supplementary fig. 1, Supplementary Material online and see Materials and Methods) identified a clade of 18 samples with consistently large (≥0.1) identity-by-state (IBS) distances from the rest of the samples, which we reasoned may represent sample swaps or human contamination. Examination of mitochondrial reads from these samples suggested in all cases that they were not of chimpanzee origin (supplementary table 3, Supplementary Material online), and we thus excluded them from any further analyses. Our final data set therefore contained 468 samples.

Population Structure and Genetic Ancestry of Transcriptome Samples

To identify the ancestry of the RNA-seq samples, we merged our genotype data with 28,559,256 SNPs genotyped in 59 wild-born chimpanzees through high coverage whole-genome sequencing (de Manuel et al. 2016). Because we were merging RNA-seq-derived variants with whole genome variants, we excluded all SNPs with missingness ≥5% (we note that our results are robust to this choice of threshold, see supplementary fig. 2, Supplementary Material online). This led to a merged data set containing only 11,679 SNPs across all 527 samples, which we confirmed was sufficient to capture subspecies differences through principal component analysis (PCA) of the wild-born samples alone (fig. 1*A* and *B*). This number of SNPs was sufficient to clearly differentiate Western and Nigeria-Cameroon chimpanzees from Central

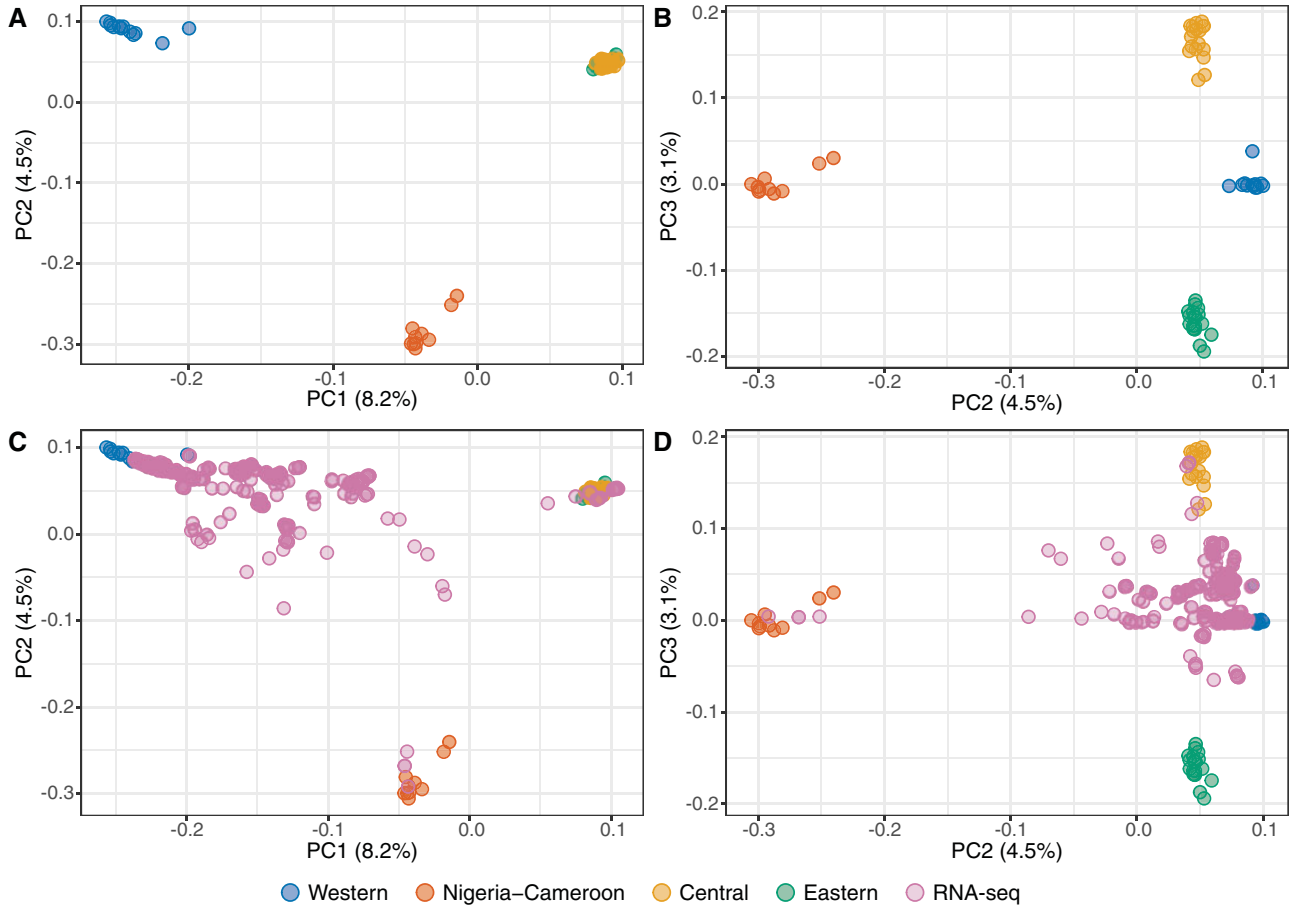

Fig. 1.—PCA of 527 chimpanzee samples with genotype data (*A*, *B*) PC1–PC2 and PC2–PC3 of 59 wild-born samples of known ancestry (de Manuel et al. 2016) using 11,679 SNPs genotyped across the entire data set. (*C*, *D*) Projection of 468 chimpanzee RNA-seq samples of unknown ancestry onto the PCA of wild-born samples. Colors indicate the four distinct chimpanzee subspecies and samples of unknown ancestry from public RNA-seq data sets.

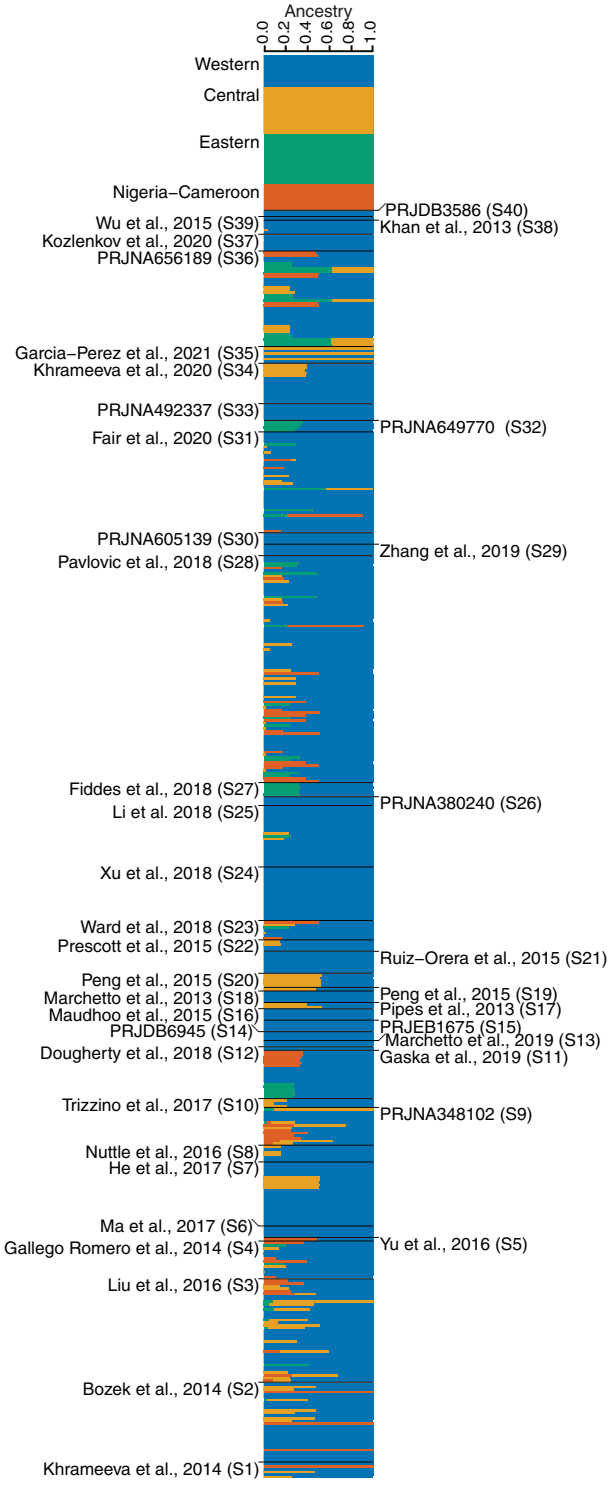

Fig. 2.—Supervised ADMIXTURE on 468 chimpanzee RNA-seq samples of unknown ancestry Samples from de Manuel et al. (2016) are shown at the top; internal study nomenclature is given in parenthesis for the 40 RNA-seq data sets with samples of unknown ancestry; full references are available in supplementary table 1, Supplementary Material online. Within a study, samples are ordered by SRA identifier.

and Eastern chimpanzees on PC1 and 2, and to further separate Central and Eastern chimpanzees from each other on PC3, recapitulating trends observed in the WGS data set.

We then projected the 468 RNA-seq samples onto this PCA (fig. 1*C* and *D*). A substantial fraction of RNA-seq samples cluster with Western chimpanzees, and a further three individuals fall within the Nigeria-Cameroon chimpanzee cluster. However, the majority of samples fall outside the four subspecies clusters, suggestive of mixed ancestries and of the presence of substantial levels of admixture in our data set. Examination of PC2 and PC3 suggested that there are four samples of possible unadmixed Central origin in the RNA-seq data, and none of unadmixed Eastern origin.

To refine these ancestry inferences, we then performed a supervised ADMIXTURE analysis with the RNA-seq samples, using the 59 known individuals as a reference (fig. 2). The results of this analysis broadly recapitulated trends observed in the PCA, including an outsized amount of Western ancestry within samples. In determining the ancestry composition of a sample, we consider it to be associated with a particular subspecies if that subspecies contributes $\geq$6.25% to a sample's genome, equivalent to having one unadmixed great-great grandparent from that subspecies. Using this definition, 452 of the 468 samples had some Western ancestry, and 287 were exclusively of Western ancestry (fig. 2 and table 1). The average proportion of Western ancestry among these 452 samples, including admixed individuals is 83.4%. An unsupervised ADMIXTURE analysis of all samples ($K = 4$) broadly recapitulates the supervised results and supports our ancestry inference approach (supplementary fig. 3, Supplementary Material online).

Using the same criteria only 51 samples showed evidence of Eastern ancestry, with the average proportion being 32.6% and with no sample being assigned exclusively to this subspecies. In the 59 samples with some Nigeria-Cameroon ancestry this value is 37.7%, and in the 90 samples with Central ancestry it is 33.8%. As suggested by our PCA, three samples are predicted to be of entirely Central ancestry, and four are exclusively of Nigeria-Cameroon ancestry. The remaining 174 samples show substantial contributions from more than one ancestry, evidence of recent hybridization between subspecies; 165 of these contain a substantial amount of Western ancestry, and 10 samples show substantial contributions from three of the subspecies, although none carry ancestry from all four. Most samples are either fully Western or likely had a Western parent or grandparent, making this the clear predominant ancestry source in functional genomics chimpanzee data. Reassuringly, with few exceptions, which we discuss below in more detail, both known replicates and duplicates showed consistent predictions across BioSamples and BioProjects, suggesting an overall low number of sample misclassifications/swaps in our data set.

Table 1
Summary of ADMIXTURE Results

	Western	Nigeria-Cameroon	Central	Eastern
Number of samples with at least 6.25% inferred ancestry:	452	59	90	51
Average inferred ancestry (%)	87.7	37.7	35.6	32.6
Max ancestry (%)	99.9	99.9	99.9	62.2

Genetically Distinct Individuals Represented in Chimpanzee RNA-Seq Data Sets

To begin examining how many different chimpanzees were represented within the 468 samples in our data, we calculated pairwise IBS distances between all samples and built a UPGMA dendogram (supplementary fig. 4, Supplementary Material online). As expected, known replicate samples clustered together. The only exception to this was a kidney sample from individual S2-c5-M-38y (naming convention for samples is "Study ID—original sample ID—sex—age;" not all of this information is available for all samples), which failed to cluster with three other replicates from the same individual. This could be a case of mislabeling, with the kidney potentially belonging to another chimpanzee. Although 296 of samples were of either neural/brain (167) or cardiac/muscle (129) origin, we observed no tissue-of-origin impact on sample clustering. Instead, different tissue samples from the same individual robustly clustered together, both within and between studies, confirming that UPGMA clustering could be used as a means of determining sample identity.

As a second layer of analysis and to further validate our replicate groupings, we calculated pairwise relatedness and IBS across all 468 samples in the data using Somalier (Pedersen et al. 2020), a tool designed to identify sample swaps across large high-throughput sequencing data sets (fig. 3*A*). We then asked how many of the 530 known replicate pairs (supplementary table 2, Supplementary Material online) had a relatedness estimate $\geq$0.5, the recommended threshold for identifying identical samples from RNA-seq data. Only four pairs failed to meet this threshold. Three pairs included the kidney sample from S2-c5-M-38y described above (supplementary fig. 4, Supplementary Material online); pairwise relatedness of this sample with all other S2-c5-M-38y replicates was <-1 (fig. 3*A*). The final pair had a relatedness value of 0.489 and included a muscle and a brain sample from the same replicate set. On the whole, relatedness estimates within sets of replicates were high, suggesting genotype sparsity and differences in tissue of origin did not broadly hinder our ability to identify identical samples. Excluding the three clearly unrelated pairs described above, mean relatedness amongst pairs within a replicate set was 0.884 (SD = 0.085), whereas mean pairwise relatedness in the whole data set was -0.211.

An additional 801 sample pairs (involving a total of 351 samples) not belonging to a replicate set also had relatedness $\geq$0.5. As the 468 samples used in these analyses included both known duplicate samples and, likely given the relatively small captive chimpanzee population, close relatives, which Somalier is not designed to detect, this finding was not surprising. On the whole, these analyses suggested that we are well powered to determine whether sets of replicate samples in our data are actually identical or not.

Given this, we then filtered our data set to retain only one sample from each replicate set (see Materials and Methods), resulting in a set of 237 samples. An UPGMA dendogram (supplementary fig. 5, Supplementary Material online) of these revealed 71 sets of duplicated samples across studies, all of which met at least one of the following criteria: clustering together with terminal node edge length $<$0.006, or clustering together with node length $\geq$0.006 but supporting metadata (age, sex) in common; only two clusters (5 and 9) fell under the latter. Although metadata support for most of these duplicate sets was strong, in some cases available information was scant or incongruous (e.g., Cluster 59, where sampling ages differ by 9 years despite both samples being from brain tissue likely collected post-mortem).

We then repeated the Somalier analysis with the set of 237 samples (fig. 3*B*). Average pairwise relatedness between the 168 sample pairs from the same cluster was 0.900 (SD = 0.089). However, two pairs, both part of Cluster 69 (supplementary fig. 5, Supplementary Material online) had relatedness $<$ 0.500. Examination of these pairs revealed that they had S37-BooBoo-M-28y in common; relatedness is 0.427 between S37-BooBoo-M-28y and two other samples in Cluster 69, and 0.803 between S37-BooBoo-M-28y and S3-c25-M-23y. In turn, S3-c25-M-23y had a pairwise relatedness of 0.944 with the other two samples in the cluster. Using combined evidence, we still assigned S37-BooBoo-M-28y to cluster 69. The mislabeled kidney sample from S2-c5-M-38y formed a cluster with S3-c16-F-6132 (cluster 57, pairwise relatedness 0.750), a brain sample from the Southwest National Primate Research Center (SNPRC) with no additional replicates.

We also noticed some likely sample swaps, where known duplicate samples failed to cluster together. For instance, four samples from S23 appeared to be mislabeled amongst themselves (supplementary figs. 4 and 5, Supplementary Material online, clusters 7, 12, 24, and 71; the authors of the study [Ward and Gilad 2019]) confirmed the sample swap occurred at the time of data upload to NBCI and has since been

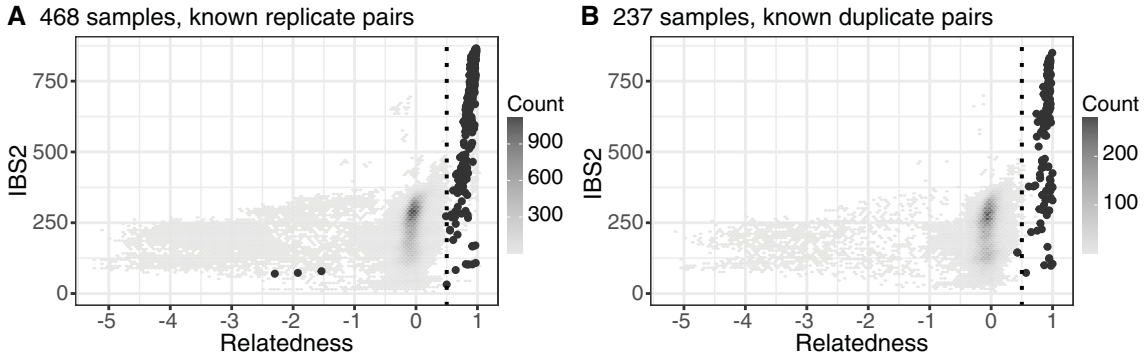

Fig. 3.—Relationships and cryptic relatedness across samples (*A*) Pairwise IBS2 (identity-by-state two; number of sites that are identical by state for a pair of samples) and relatedness, as estimated by Pedersen et al. (2020), in the unfiltered set of 468 samples. Known replicate pairs are colored dark gray. (*B*) as (*A*), but only considering 237 nonreplicate samples. Known duplicate pairs are colored dark gray.

corrected; Ward M to IGR, personal communication), whereas S18-PR01209-M-2y clustered with cell line PR00818 (cluster 33) rather than with other instances of PR01209 (cluster 28). Cluster 14 contains three samples with different names and metadata, S27-S008919-F-10y, S31-724-F-15y, and S28-4933-F-6y, all with pairwise relatedness > 0.9. Searching in Cellosaurus (Bairoch 2018) revealed that S008919 is a cell line that was part of the Yerkes National Primate Research Center collection available from the Coriell Institute, and which according to a partial Yerkes pedigree shared with us by B. Pavlovic, B. Fair, and Y. Gilad was established from a Yerkes chimpanzee with internal sample ID 724, confirming they are the same animal. Conversely, relatedness between S28-4933-F-6y and S11-S4933-F-6y, which should be the same individual, was 0.206. In addition, relatedness was high between S28-4933-F-6y and S28-495-M-NA (0.505) or S31-495-M-NA (0.517), which are duplicate instances of Amos, the father of S008919. We thus concluded that S28-4933-F-6y is mislabeled, and actually derived from S008919.

Identifying Cryptic Relatives in Public Chimpanzee RNA-Seq Data Sets

In our Somalier analyses, 18 pairs of samples not assigned to the same duplicate cluster had relatedness > 0.5 (supplementary table 4, Supplementary Material online and supplementary fig. 6, Supplementary Material online), again hinting at the presence of cryptic relative pairs in the data. Thus we filtered our data to a final set of 135 genetically unique individuals (see Materials and sMethods), and computed pairwise identity-by-descent (IBD) metrics for these individuals. Pairwise Z0, Z1 and Z2 values describe the genome-wide probability that at a given site a pair of samples will share 0, 1, or 2 identical-by-descent alleles, respectively. Blay et al. previously tested the validity of this approach to detect kinship in human data sets with known family relationships. They found that as expected, parent-offspring pairs shared 1 IBD allele at most sites ($Z1 \sim 1$). Similarly, siblings typically shared between 0 and 2 IBD alleles and second-degree relatives shared 1 IBD allele at half the sites ($Z1 \sim 0.5$), and that these inferences can be made robustly from SNPs ascertained from RNA-seq data (Blay et al. 2019).

Using these values as a guide, we took advantage of known relationships in our data set to define thresholds specific to our data. We first analyzed only the 38 samples from S31, which contains eight known pairs of first-degree (all parent/offspring) relatives (Fair et al. 2020). All eight of the known pairs had IDB $Z1 > 0.60$, and seven had IBD $Z1 > 0.70$. In addition, we were aware of fair pairs of known second-degree relatives; IBD Z1 scores for these pairs ranged from 0.19 to 0.60 (fig. 4*A*). We used these results, alongside previous observations from Blay et al. (2019), to define conservative IBD Z1 and Z2 thresholds that allowed us to identify possible additional relative pairs in the rest of our data.

We applied the same approach to all research centers or zoos with more than five samples, as well as to the entire data set (fig. 4*B–F*, table 2, and supplementary table 5, Supplementary Material online). We identified only 27 pairs of first-degree relatives across the entire data set, but 146 likely second-degree relative pairs, again suggesting that the overall captive chimpanzee population is small. Of note, 97 of the likely second-degree relative pairs involved individuals sampled in different sites, highlighting connections between them. Within the Yerkes data set (fig. 4*C*), we identified a pair of samples with IBD Z2 0.7455, which we predicted to be derived from the same donor individual. This assignment is partially supported by incomplete metadata for the two individuals; there are two additional cases in the full data set (fig. 4*B*; IBD $Z2 > 0.65$ in all three cases). Below we have retained both members of all three pairs, but these observations suggest that the true number of genetically unique individuals in our data set is between 135 and 132.

Our analyses also readily identified four clear full sibling pairs, visible in the center of figure 4*B*, in the full data set. Three of these involved the same individual, S28-462-F-42y, and three

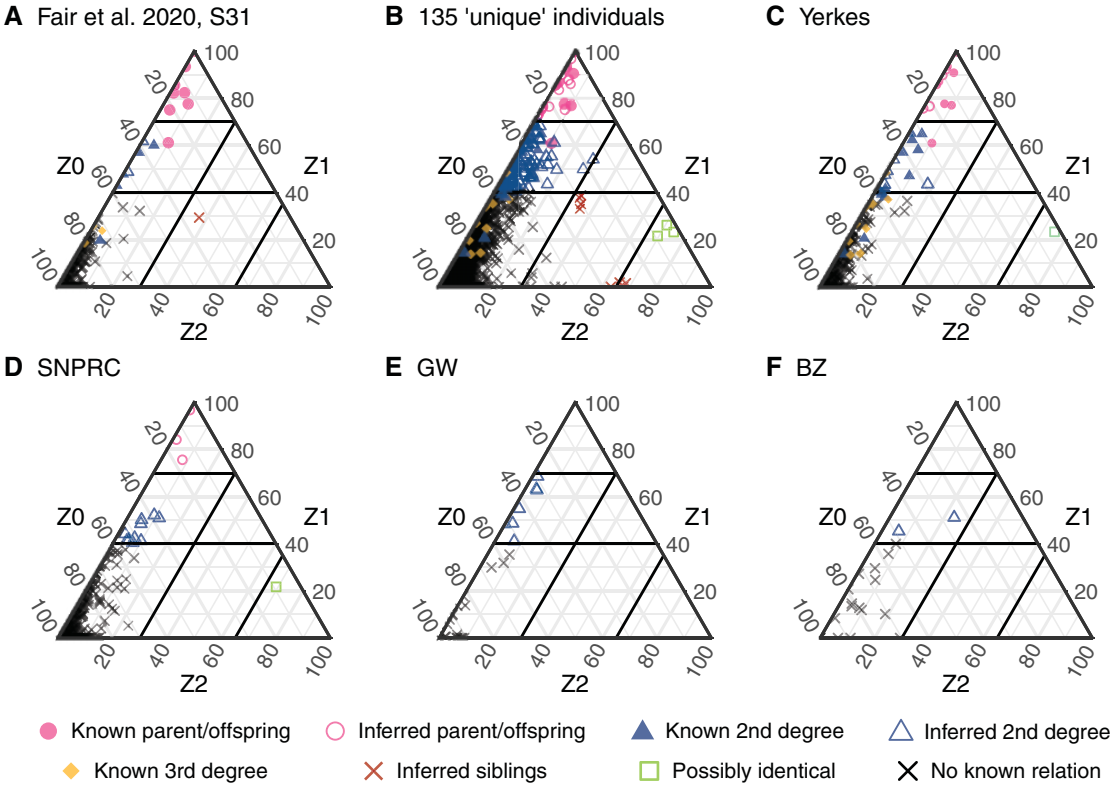

FIG. 4.—Predicted relatedness amongst 135 chimpanzee samples in public RNA-seq data Ternary diagrams showing pairwise IBD estimates across different data subsets. Thresholds for defining different relationships between pairs are indicated by black lines on the diagram. Yerkes: Yerkes National Primate Research Center; SNPRC: Southwest National Primate Research Center; GW: George Washington University National Chimpanzee Brain Resource; BZ: Burger's Zoo in Arnhem.

Table 2
Relatedness Summary across Sampling Sites

	All Samples ($n = 135$)	Fair et al. ($n = 38$)	Yerkes ($n = 36$)	SNPRC ($n = 36$)	GW ($n = 12$)	BZ ($n = 9$)
Possibly identical pairs	3	0	1	1	0	0
Parent/offspring pairs	27	9	14	3	0	0
Second-degree pairs	146	8	22	10	7	2
Full siblings pairs	9	1	0	0	0	0

other animals, which should therefore also be full siblings with one another. Instead, these additional pairs involving these three chimpanzees (S3-c37-M-35y, S31-4x0354-M-21y, and S36-4x0421-F-NA; Z2 between the latter two is high enough to suggest they might be the same individual) consistently showed unexpectedly high IBD Z2 values (>0.60) and a nearly complete lack of differing heterozygous sites (IBD Z1 < 0.018 in all cases), and are visible in the bottom right of figure 4*B*; a fifth chimpanzee, S35-CH114-NA-NA also exhibited high IBD Z2 values with two of the animals in this group. Intriguingly, these individuals were sampled across a variety of centers and tissues, with no obvious links between them. Despite the lack of known full sibling pairs to guide our inferences, we cautiously labeled these individuals as siblings, whereas acknowledging unresolved complexity in their relationships.

Identifying Unique and Unrelated Individuals in Transcriptomic Data Sets

Our analyses identified between 132 and 135 genetically unique individuals in our data set; a simple UPGMA dendogram of all 135 is presented in figure 5. As may be expected, we observed significant reuse of samples across studies (mean number of studies an individual appears in = 1.7), with the two most frequent individuals in our final data set being cell lines: PR00818 appeared in nine BioProjects and C3649 in five. A further 53 samples appeared twice, 12 thrice and 4 four times. On average, 1.35 tissues have been studied per animal, with the brain and the heart being the most commonly sequenced ones—across 58 different animals in the case of brain or neural tissue and 60 in that of cardiac or muscle tissue. Notably, these are broadly distinct sets of

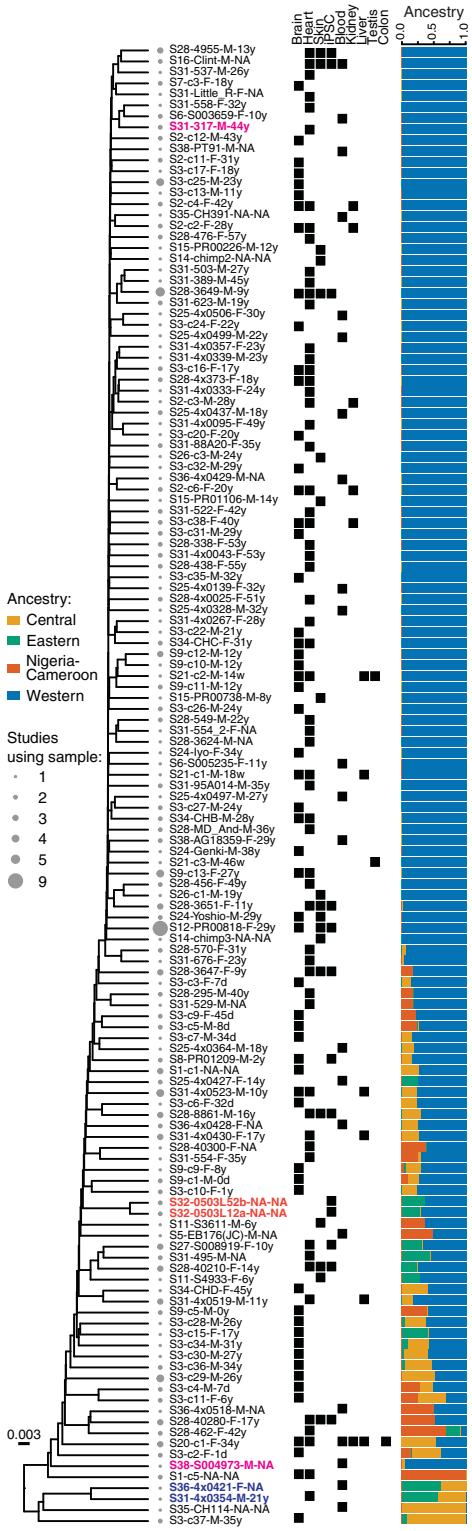

Fig. 5.—A total of 135 genetically distinct chimpanzee individuals in public RNA-seq data sets UPGMA dendogram of 135 samples predicted to be genetically distinct in our data set. The three sample pairs predicted to be closely related by relatedness and IBD analyses are indicated by colored taxon labels. Filled boxes indicate tissues sampled across studies from each individual. Ancestry estimates are as in figure 2.

individuals; animals studied for their brain tended to not be considered much further (mean number of tissues sequenced = 1.29), and the same is true of cardiac samples (mean number of tissues sequenced = 1.13).

As our analyses above demonstrated, these individuals are not unrelated, nor do they represent a random sample of the genetic diversity of the wild chimpanzee population. Beyond the three sample pairs that appeared to be genetically identical, 72 individuals had at least one second-degree relationship with another individual in the data set, and 36 were part of at least one parent/offspring or sibling pair; a single individual was part of 4. On average, any individual had at least 2.09 second-degree relatives in the data set, and 7 had over 10; in the most extreme instance we inferred 19 second-degree relationships involving S3-c34-M-31y. We iteratively excluded individuals with the most relationships until no first or second-degree relative pairs remained in the data set; this subset contained between 86 and 89 individuals, depending on which relative we removed from certain pairs.

Furthermore, of the 135 individuals, 85 (63%) were entirely of Western ancestry (>93.75% as predicted by ADMIXTURE), and 130 out of 135 (96%) have significant (>6.25%) Western chimpanzee ancestry. There were 48 hybrid individuals, 42 of which are of two different ancestries and 6 of three different ancestries. The final data set contained only one Central chimpanzee, one Nigeria-Cameroon chimpanzee, and no Eastern chimpanzees at all.

Genetic diversity, tissue type, and relationships amongst the 15/14 chimpanzee iPSC lines in our data set. Ancestry estimates are as in figures 2 and 5. Samples that appear to be mislabeled are indicated by the red box next to individual sample IDs, multiple samples predicted to be from the same genetically unique individual are marked in the final ID column.

Finally, we considered the subset of samples from iPSCs or iPSC-derived cell types (fig. 6). As iPSCs can be indefinitely maintained and experimentally perturbed, and, upon directed differentiation, give access to a large number of otherwise unobservable tissue types and cell states, they have recently become established as a model system in comparative functional genomics studies (Dannemann and Gallego Romero 2021). However, because they are time-consuming to establish, only 15 (possibly 14, depending on how we resolve the high similarity between S32-0503L12-NA-NA and S32-0503L52-NA-NA) have been established to date; each of them has been included in an average of 2.6 studies.

Of the 15 different individuals in our data with iPSC data, 6 are entirely Western and the rest are two-ancestry hybrids with at least one Western parent, again highlighting the difficulty of successfully capturing existing chimpanzee genetic diversity in functional genomics resources. Additionally, our IBD analyses suggested that C40210 and C8861 are second-degree relatives, as are C40210 and S008919. Notably, all but one of the sample swaps we described above involve iPSCs or

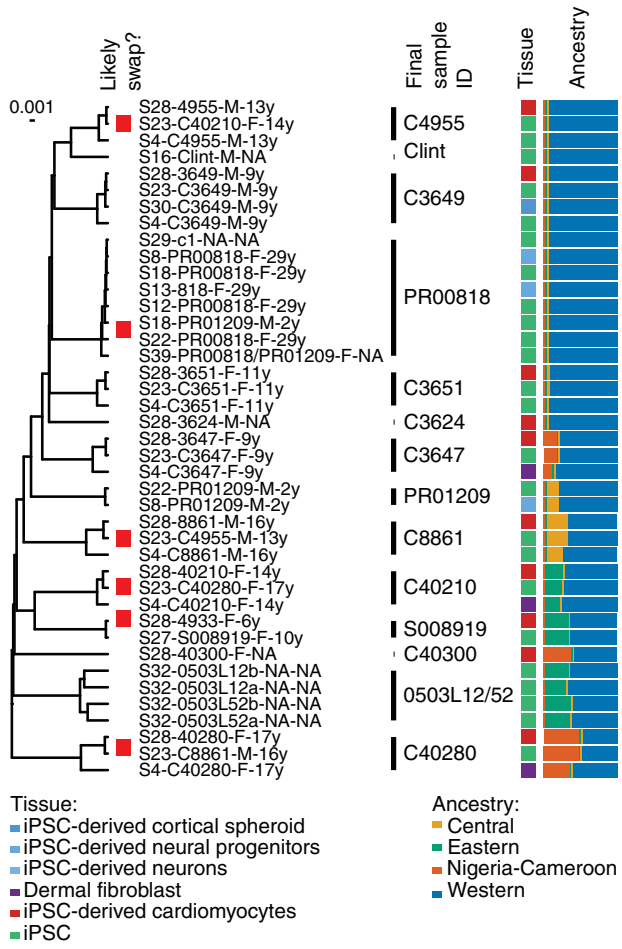

Fig. 6.—Genetic diversity and sample swaps amongst iPSC and iPSC-associated samples

iPSC-derived samples, emphasizing the increased complexity of cell culture relative to observational data generation from frozen tissues (Chatterjee 2007).

Discussion

Comparisons with chimpanzees are essential for understanding human evolution and the genetic basis of human-specific traits. However, chimpanzees are a critically endangered species, and increasing their captive population is unethical, and indeed illegal in many countries. Samples are therefore strictly limited to those from animals currently living in captivity, and to post-mortem tissue and cell line collections. These populations were established decades ago, sometimes with little regard to subspecies differences, and have not always been managed to maximize genetic diversity. Breeding practices in captive colonies, especially prior to the establishment of genetically guided breeding programs, have produced hybrids with highly variable ancestry components and this lack of specificity makes them a poor resource (Hvilsom et al. 2013).

Our meta-analysis of chimpanzee individuals in transcriptomics provides important information for contextualizing comparative research done so far, especially as many studies do not include more distantly related outgroups. We consistently find that both in the number of samples and proportion of ancestry there is an overwhelming bias toward Western chimpanzees (*Pan troglodytes verus*) in public data sets. This is particularly noteworthy as, of the four extant chimpanzee subspecies, Western chimpanzees have the smallest effective population size, and show the highest amount of genetic drift (Prado-Martinez et al. 2013; Lester et al. 2021). All 40 studies we considered included at least one unadmixed individual of Western ancestry. In contrast, only two studies sampled a Nigeria-Cameroon chimpanzee (as we later discovered, this was the same individual); and only S35 (García-Pérez et al. 2021) sampled a purely Central chimpanzee. None of the studies considered subspecies when sampling.

We also observe ample evidence of cryptic relationships, further pointing toward a depletion of genetic diversity. We are able to identify a total of 135 (possibly 132) unique individuals and a maximum of 89 unrelated individuals within our initial set of 468 chimpanzee RNA-seq samples. At least 82 samples have a first or second-degree relationship with another sample in the data set. As expected from captive-bred individuals, some of these relationships are between samples from the same source. However, we see a significant number of cryptic relationships between institutes. This means that even sampling from multiple sources or from tissue repositories like GW could include some cryptically related samples.

In addition, many individuals in our data have been sampled more than once. The least concerning scenario of repeat sampling is sequencing of different tissues from the same donor—but only 29 individuals have transcriptomic data from more than one tissue type. If we consider the 58 individuals with neural RNA-seq data, 37 appear in more than one study, and 21 of those only in the context of neural samples across different studies. As half of these samples are derived from the National Chimpanzee Brain Resource at George Washington University, a post-mortem brain bank, it is likely that these studies are repeatedly sequencing the same brain sample. Similarly, of the 60 heart/muscle samples, 40 have been used more than once, 17 only for heart/muscle purposes—all 17 have been used by both S31 (Fair et al. 2020) and S28 (Pavlovic et al. 2018), which were conducted by the same research group using partly overlapping collections of individuals/cell lines. This finding has implications about how much unique transcriptomic information is actually available.

Unlike with human studies, however, more diverse sampling of chimpanzees is largely not possible. The only currently viable solution is deliberate sampling of captive donors of specific subspecies, and the expansion of existing iPSC collections. Unfortunately, genetic surveys of captive chimpanzees in zoos and research institutes worldwide consistently find

that over 75% of captive individuals are of solely Western ancestry (Ely et al. 2005; Hvilsom et al. 2013; Carlsen and de Jongh 2014), and indeed, they are the only subspecies for which a genetically guided breeding program has been established, by the European Association of Zoos and Aquaria (Carlsen and de Jongh 2014). The large number of hybrids in our data set—48 of the 135 true individuals—is also in accordance with captive colony statistics, as 531 of the 1059 individuals in the European Chimpanzee Studbook are hybrids (Carlsen and de Jongh 2014).

Because so little is known about the variability of gene expression between chimpanzee subspecies, it is difficult to quantify the degree to which the poor representation of chimpanzee genetic diversity hinders inter-species comparisons. However, multiple lines of evidence argue that poor sampling is likely to lead to incomplete inferences. Although they had modest sample sizes and featured primarily individuals of Western ancestry, the two largest studies of gene expression in chimpanzees, Pavlovic et al. (2018) and Fair et al. (2020), have both shown that inter-individual differences in gene expression have a clear genetic component in both humans and chimpanzees. In parallel, genomic studies of all four chimpanzee subspecies have identified multiple highly differentiated loci that show strong evidence of hard adaptive sweeps in response to the local environment, especially in immune-associated genes in Eastern chimpanzees (Schmidt et al. 2019), which are the worst represented in our data.

Other organisms offer additional evidence. In mice, where the use of a single inbred strain in a study is common, the same local genotype can lead to very different phenotypes across strains and genetic backgrounds (Sittig et al. 2016; Li and Auwerx 2020). In humans, which contain an order of magnitude less genetic diversity than chimpanzees, both genome-wide association studies and expression quantitative trait loci mapping studies have systematically identified loci that have population-specific effects (Daly 2010; Martin et al. 2019; Sirugo et al. 2019), and the systematic overrepresentation of individuals of European ancestry has drastic implications for the generalization of medical genetic research and the extent to which its benefits can be translated into other populations. Thus, if not accounting for diversity in human populations can have substantial limitations on the information that can be learnt from them, not accounting for the significantly larger amount of variation found across chimpanzee subspecies when doing comparative research doubtlessly confounds attempts to truly understand humans as great apes.

Materials and Methods

Sample Selection

As of the 19th of January 2021, the NCBI SRA database contained 4,389 RNA-seq samples with taxon ID *Pan troglodytes*, sequenced through 17 different technologies, with the main ones being the Illumina HiSeq 2000 (483 samples), Illumina HiSeq 2500 (2,316 samples), and HiSeq 4000 instruments (422 samples). To avoid potential biases from calling SNPs on sequencing data generated with vastly different technologies (Hwang et al. 2015), we considered only samples sequenced on these three instruments, retaining 3221 samples across 49 distinct BioProjects (collections of BioSamples from a single initiative in NCBI) for further consideration. Four single-cell RNA-seq studies (Mora-Bermúdez et al. 2016; Kronenberg et al. 2018; Kanton et al. 2019; Pollen et al. 2019) accounted for 2,164 of these samples, which we excluded from further analysis as our genotyping pipeline yields low quality genotype calls from single-cell data due to read scarcity.

The remaining 1,059 entries were associated with a total of 808 BioSample IDs; we randomly chose one entry per BioProject for further processing. For studies with more than 5–6 replicates per individual (as identified by metadata on SRA), we only retained five randomly chosen replicates in order to reduce computational load; thus our description of sample swaps and mislabeled samples is limited. For paired-ended samples, we retained only the R1 file. Two data sets (PRJNA445737 and PRJDB1766) failed processing with GATK. We additionally excluded two data sets (PRJNA385016, PRJNA481380) as the individuals in these studies were already well-represented in our data set. Our final data set includes 486 BioSamples, 339 of which were single ended, and 147 pair ended, from 40 different BioProjects (supplementary tables 1 and 2, Supplementary Material online) and spans 20 different tissue or cell types (and two samples labeled "pooled tissues"); for simplicity, when plotting we collapsed these tissue/cell types into nine different categories:

1. Brain: brain, iPSC-derived cortical spheroids, iPSC-derived neural progenitors, iPSC derived neurons
2. Heart: heart, iPSC-derived cardiomyocytes, muscle, myoblast
3. Blood: whole blood, peripheral blood mononuclear cells, lymphoblastoid cell lines, endothelial tissue
4. Skin: skin, dermal fibroblasts
5. iPSCs
6. Kidney
7. Liver
8. Testis
9. Colon

In parallel, we downloaded the complete set of 28,559,256 SNPs genotyped in 59 wild-born chimpanzees from all four subspecies (12 Western, 19 Eastern, 10 Central and 19 Nigeria-Cameroon) generated by de Manuel et al. through high coverage whole-genome sequencing (de Manuel et al. 2016). We used CrossMap 0.3.8 (Zhao et al. 2014) to convert

genomic coordinates for these SNPs from panTro4 to panTro5 using chain files from the UCSC Genome Browser. A total of 1,626 unclassified contigs were then removed with BCFtools 1.9 (Li et al. 2009). Fewer than 1% of SNPs mapped to these contigs.

SNP Calling

We removed all unclassified contigs from the panTro5 genome release. Then we implemented the GATK 3 version of GATK's RNAseq short variant discovery (SNPs + Indels) pipeline (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-, last accessed November 18, 2021). We additionally incorporated an initial quality control step to this process, and used Trimmomatic 0.38 (Bolger et al. 2014) to remove reads with average quality below 20 and minimum length below 40 bp, as some studies had read lengths of 50 bp. We applied Trimmomatic separately on longer samples and ensured that this relaxed filter did not have a sizeable impact on the number of reads kept after trimming.

To determine values for variant calling filters, we considered existing literature (Prado-Martinez et al. 2013) and GATK recommendations. Using a test run of 15 samples, we implemented to following filters, which produced enough high quality SNPs for downstream analyses: Fisher Strand (FS) < 26; total depth (DP) < 10 (especially relaxed as sequencing depth varies considerably between studies and between genes); mapping quality (MQ) > 25; genotype quality (GQ) > 20.

Merging and Filtering

A total of 7,529,801 variants passed the variant calling filters above. We then used VCFtools 0.1.15 (Danecek et al. 2011) to remove indels and variants that had failed variant filtering criteria described above, for a final set of 6,943,957 SNPs. We used PLINK 1.90 (Purcell et al. 2007) to calculate a site frequency spectrum for retained SNPs, which suggested widespread presence of duplicate samples (supplementary fig. 7, Supplementary Material online). As expected, the majority of variants were singletons, which we also removed with VCFtools.

We then used BCFtools to merge our genotype calls with the reference WGS SNPs, and applied different missingness thresholds of 2%, 5%, and 10% to both the unmerged and merged data sets. After considering results obtained at these different thresholds (supplementary fig. 2, Supplementary Material online and supplementary table 6, Supplementary Material online), we elected to proceed with a missingness threshold of 5%, which allowed us to account for the effect of tissue-specific and temporal expression patterns that characterize the different RNA-seq samples. This resulted in a final set of 54,706 SNPs genotyped in our samples, as well as 11,659 SNPs in our samples and the WGS data.

Relatedness Analyses

We used PLINK to compute genome-wide IBS pairwise distances between all samples. We then subtracted the IBS values from 1 to get a distance matrix. Using the *ape* package (ver 5.4-1, Paradis and Schliep 2019) in R 4.0.5 (R Core Team 2020) we then performed hierarchical clustering on these and plotted dendograms using treeio 2.2.4 (Wang et al. 2020). As the evolutionary distance between our samples is not consequential, we arbitrarily chose the UPGMA method for dendogram generation. In parallel, we made use of Somalier 0.2.10 (Pedersen et al. 2020), which takes polymorphic sites and calculates pairwise relatedness (coefficient of relationship, defined as $(\text{shared-hets}_{i;\,j} - 2*\text{ibs0}_{i;\,j})/(\min(\text{hets}_i; \text{hets}_j))$ between each pair of samples. Approximately 1,000 SNPs with a minor allele frequency as close to 0.5 as possible are needed for a high-confidence analysis; we therefore considered only the 946 SNPs in our data with MAF ≥ 0.25. When filtering replicate or duplicate samples from clusters, we chose at random for sets that contained 2 individuals, and retained the one with the highest average relatedness to the rest of individuals in the cluster otherwise.

Because neither set of results provided enough resolution to detect higher-degree relationships, we also used PLINK to calculate pairwise IBD metrics Z0, Z1, and Z2 to identify relative pairs within our data set as in Blay et al. (2019). We defined the following thresholds for identifying relatedness between pairs on the basis of known pedigree relationships, metadata, and (Blay et al. 2019): parent-offspring: $Z1 \geq 0.7$; inferred siblings: $Z2 \geq 0.30$, $Z1 \geq 0.30$ and $Z0 \leq 0.40$; second-degree relatives: $04 \leq Z1 < 0.7$; potential identical: $Z2 \geq 0.65$ and $Z0 < 0.10$. Ternary diagrams were generated with the ggtern3.3.0 (Hamilton and Ferry 2018) package.

Population Structure and Ancestry Assignments

We used SmartPCA (part of Eigensoft 7.2.1) (Price et al. 2006) to perform PCA in the 59 wild-born samples and confirm that it was possible to recover population structure across the species with low numbers of SNPs. Our results at various missingness thresholds were comparable to those observed when using genome-wide data. We then projected our 486 RNA-seq samples onto this space, again using SmartPCA.

We also used ADMIXTURE 1.3.0 (Alexander et al. 2009) to quantify subspecies ancestry of the RNA-seq samples. We set $K = 4$, hoping to identify the four chimpanzee subspecies, but the presence of duplicates and significant admixture in the RNA-seq samples (both known and unknown) confounded the algorithm at $K = 4$ because identical individuals sampled repeatedly clustered together as separate populations. Therefore, we used the WGS samples as reference to perform a supervised ADMIXTURE analysis. RNA-seq samples were assigned a particular subspecies ancestry if that subspecies contributed more than 6.25% to its ancestry, equivalent to an unadmixed great-grandparent.

Identification of Ambiguous Samples

Our initial dendogram of 486 samples included a clade of 18 samples that had consistently large pairwise IBS distances (>0.1) with all other samples in the data (supplementary fig. 1, Supplementary Material online). Reasoning that these might represent sample swaps or contamination issues, we mapped and quantified raw reads using Kallisto 0.45.1 (Bray et al. 2016) against the mitochondrial genomes of chimpanzee (RefSeq NC_001643.1), human (rCRS, NC_012920.1), orang-utan (NC_002083.1), and rhesus macaque (NC_005943.1), as these were the other taxa included in the studies associated with these samples. In all cases, we identified a significant fraction of reads originating outside chimpanzees, confirming our intuition (supplementary table 3, Supplementary Material online); other samples from our data set that fell within the main part of the dendogram did not exhibit this pattern. We therefore removed these samples from all downstream analyses.

Analysis Code

All analyses described were carried out using custom bash and R scripts, and are available at https://gitlab.unimelb.edu.au/navyas/chimpanzee-snp-calling (last accessed November 18, 2021).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Data Availability

All data sets in this study are publicly available. A VCF of genotype data from wild-born chimpanzees used as reference is available at https://www.biologiaevolutiva.org/tmarques/data (last accessed November 18, 2021). GEO IDs of all analyzed data sets are available in supplementary tables 1 and 2, Supplementary Material online. A VCF file containing genotype calls from all 468 RNA-seq samples, unfiltered for missingness, is available for download at Figshare with https://melbourne.figshare.com/articles/dataset/RNA-seq_derived_chimpanzee_genotypes/14822289 (last accessed November 18, 2021).

Acknowledgments

We thank Davis McCarthy, Rob Lanfear, Phillip Bayer, and members of the Gallego Romero lab at the University of Melbourne for discussion. We also thank Yoav Gilad, Bryan Pavolvic, and Benjamin Fair for sharing a partial pedigree of Yerkes chimpanzees. Michelle Ward confirmed the sample swaps affecting S23; this mislabeling has now been corrected on the SRA record.

Author Contributions

N.S. and I.G.R. designed the study, performed all analyses, and wrote the manuscript. B.S. implemented the GATK pipeline and provided comments on the manuscript.

Literature Cited

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655–1664.
Bairoch A. 2018. The Cellosaurus, a cell-line knowledge resource. J Biomol Tech. 29(2):25–38.
Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M. 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. Mol Biol Evol. 28(1):615–623.
Blake LE, et al. 2020. A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. 30(2):250–262.
Blay N, et al. 2019. Assessment of kinship detection using RNA-seq data. Nucleic Acids Res. 47(21):e136.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478(7369):343–348.
Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34(5):525–527.
Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153–1157.
Cardoso-Moreira M, et al. 2019. Gene expression across mammalian organ development. Nature 571(7766):505–509.
Carlsen F, de Jongh T. 2014. European Studbook for the chimpanzee Pan Troglodytes, 1st edition of joint EEP studbook 2014. 1st ed. EAZA, EEP. Copenhagen, Denmark: Copenhagen Zoo. Available from: https://www.zoo.dk/files/stambog_chimpanser_zoo_2014.pdf.
Chatterjee R. 2007. Cases of mistaken identity. Science 315(5814):928–931.
Coop G, Przeworski M. 2007. An evolutionary view of human recombination. Nat Rev Genet. 88(1):23–34.
Daly AK. 2010. Genome-wide association studies in pharmacogenomics. Nat Rev Genet. 11(4):241–246.
Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.
Dannemann M, Gallego Romero I. 2021. Harnessing pluripotent stem cells as models to decipher human evolution. Febs J. 1–19. doi: 10.1111/febs.15885.
de Manuel M, et al. 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354(6311):477–481.
Ely JJ, et al. 2005. Subspecies composition and founder contribution of the captive U.S. chimpanzee (*Pan troglodytes*) population. Am J Primatol. 67(2):223–241.
Fair BJ, et al. 2020. Gene expression variability in human and chimpanzee populations share common determinants. eLife 9:e59929.
Fischer A, Pollack J, Thalmann O, Nickel B, Pääbo S. 2006. Demographic history and genetic differentiation in apes. Curr Biol. 16(11):1133–1138.
Gallego Romero I, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 13(7):505–516.
García-Pérez R, et al. 2021. Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. Nat Commun. 12(1):3116.
Gonder K, Disotell T. 2006. Contrasting phylogeographic histories of chimpanzees in Nigeria and Cameroon: a multi-locus genetic analysis. In: Lehman, SM, Fleagle, JG, editors. Primate biogeography: progress and

Prospects. Boston (MA): Springer US. p. 135–168. doi: 10.1007/0-387-31710-4_5.
Hamilton NE, Ferry M. 2018. ggtern: ternary diagrams using ggplot2. J Stat Soft. 87(Code Snippet 3):1–17.
He Z, et al. 2017. Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. Nat Neurosci. 20(6):886–895.
Hvilsom C, et al. 2013. Understanding geographic origins and history of admixture among chimpanzees in European zoos, with implications for future breeding programmes. Heredity 110(6):586–593.
Hwang S, Kim E, Lee I, Marcotte EM. 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 5:17875.
Kaessmann H, Wiebe V, Pääbo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. Science 286(5442):1159.
Kanton S, et al. 2019. Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574(7778):418–422.
Kodama Y, Shumway M, Leinonen R., International Nucleotide Sequence Database Collaboration. 2011. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40(Database issue):D54–D56.
Kronenberg ZN, et al. 2018. High-resolution comparative analysis of great ape genomes. Science 360(6393):eaar6343.
Lester JD, et al. 2021. Recent genetic connectivity and clinal variation in chimpanzees. Commun Biol. 4(1):11.
Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li H, Auwerx J. 2020. Mouse systems genetics as a prelude to precision medicine. Trends Genet. 36(4):259–272.
Martin AR, et al. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 51(4):584–591.
Mora-Bermúdez F, et al. 2016. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. eLife 5:e18683.
Mostajo-Radji MA, Schmitz MT, Montoya ST, Pollen AA. 2020. Reverse engineering human brain evolution using organoid models. Brain Res. 1729:146582.
Oates JF, Groves CP, Jenkins PD. 2009. The type locality of *Pan troglodytes vellerosus* (Gray, 1862), and implications for the nomenclature of West African chimpanzees. Primates 5050(1):78–80.
Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35(3):526–528.
Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. 2018. A comparative assessment of human and chimpanzee iPSC-derived cardiomyocytes with primary heart tissues. Sci Rep. 8(1):15312.
Pedersen BS, et al. 2020. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. 12(1):62.
Pollen AA, et al. 2019. Establishing cerebral organoids as models of human-specific brain evolution. Cell 176(4):743–756.e17.
Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. Nature 499(7459):471–475.
Price AL, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8):904–909.
Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.
R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/. Accessed November 18, 2021.
Schmidt JM, M de M, Marques-Bonet T, Castellano S, Andrés AM. 2019. The impact of genetic adaptation on chimpanzee subspecies differentiation. PLoS Genet. 15(11):e1008485.
Sirugo G, Williams SM, Tishkoff SA. 2019. The missing diversity in human genetic studies. Cell 177(1):26–31.
Sittig LJ, et al. 2016. Genetic background limits generalizability of genotype-phenotype relationships. Neuron 91(6):1253–1259.
Varki A, Altheide TK. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res. 15(12):1746–1758.
Wang L-G, et al. 2020. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol Biol Evol. 37(2):599–603.
Ward MC, Gilad Y. 2019. A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from humans and chimpanzees. eLife 8:e42374.
Waterson RH, Lander ES, Wilson RK. 2005. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69–87.
Won Y-J, Hey J. 2004. Divergence population genetics of chimpanzees. Mol Biol Evol. 1322(2):297–307.
Zhao H, et al. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics 30(7):1006–1007.

Associate editor: Selene Fernández Valverde