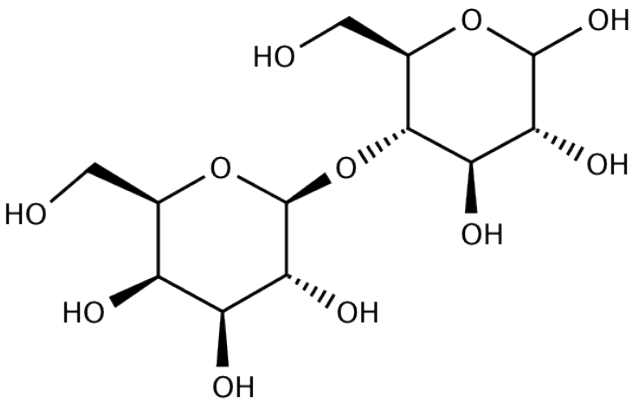

Mapping out African genomics research with 'sf '
Kirsty Lee Garson
EXEGETIC
University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad
SPES BONA

What is the Human Genome?

The collection of all of our genes,
the biological material
which shapes much of who we are

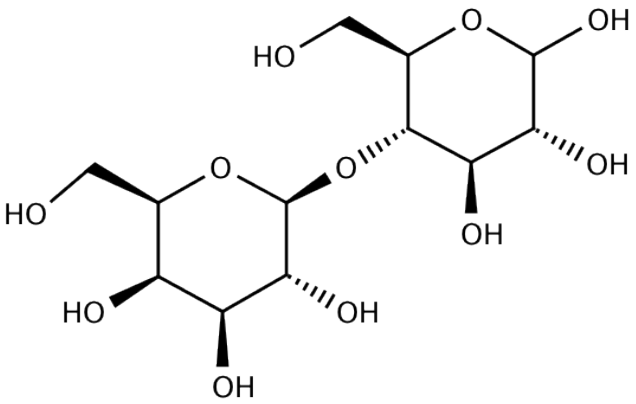
Our genes influence traits as diverse as:

- whether we can digest lactose

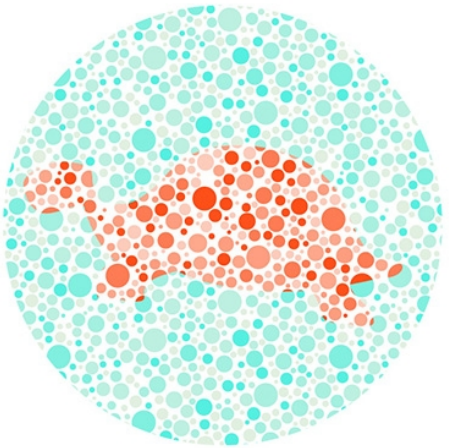


Our genes influence traits as diverse as:

- whether we can digest lactose

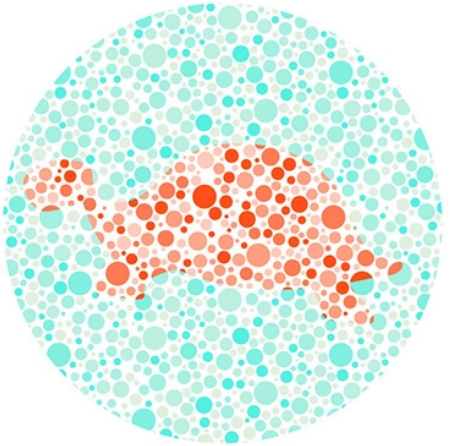


- our ability to distinguish pink from blue

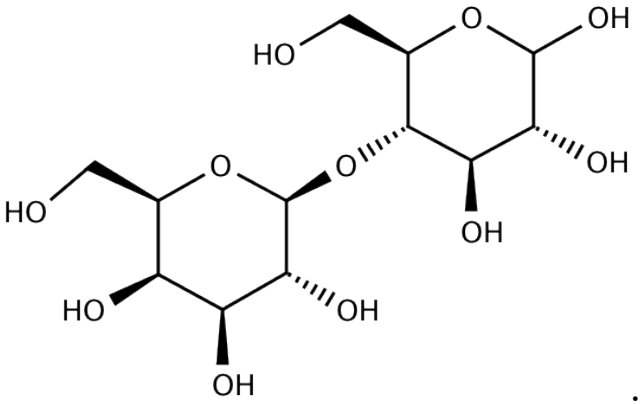


Our genes influence traits as diverse as:

- our ability to distinguish pink from blue

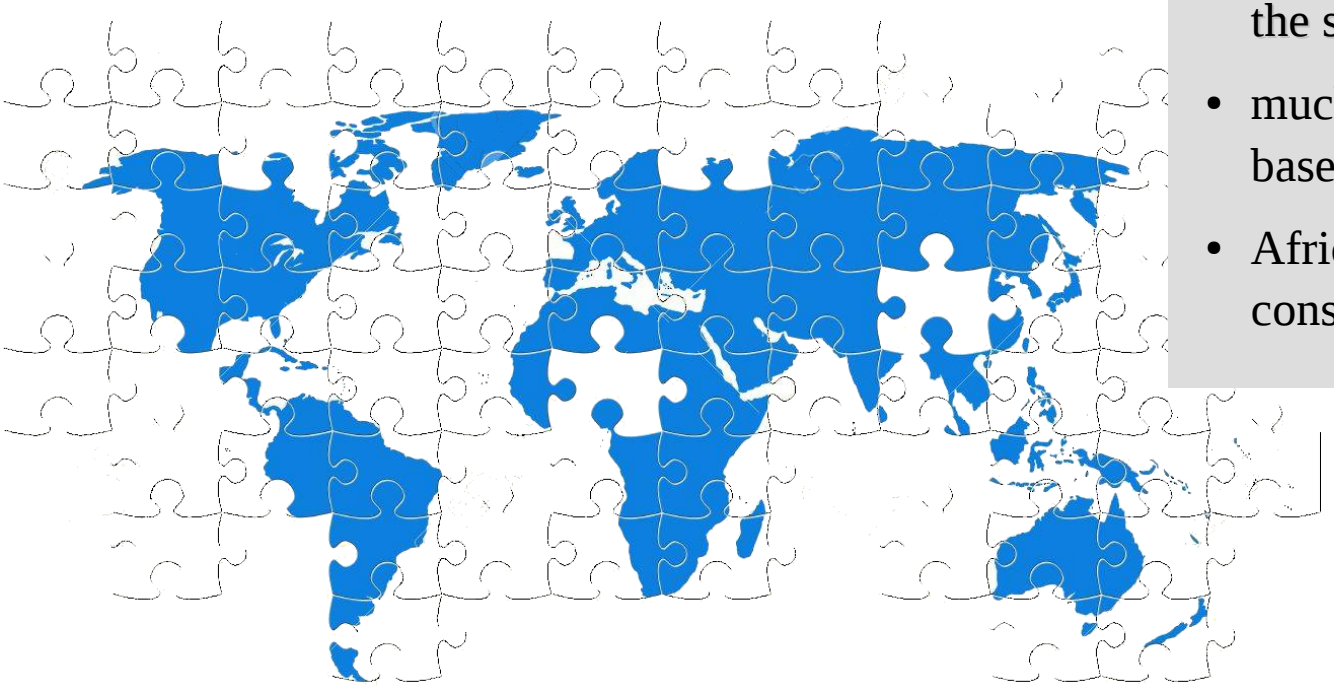


- whether we can digest lactose



- our risk of developing heart disease

What makes African genome data so valuable?



- we've only just scratched the surface
- much of what we know is based on European populations
- African populations harbour considerable genetic diversity

Recent studies exploring the human genome in African populations

VISUALIZING PROGRESS USING 'SF'

```r
# load multiple packages in a single command using 'easypackages'
library(easypackages)
packages("sf", "dplyr", "ggplot2")


# read in world borders data set (http://thematicmapping.org/downloads/world_borders.php)
map_data <- st_read(inDir, 'TM_WORLD_BORDERS-0.3.shp')


# subset african data with a bit of help from dplyr
map_data <- map_data %>%
                    filter(REGION==2)
```

```
Simple feature collection with 6 features and 11 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -8.667223 ymin: -18.01639 xmax: 31.30278 ymax: 37.09139
epsg (SRID):    4326
proj4string:    +proj=longlat +datum=WGS84 +no_defs
  FIPS ISO2 ISO3  UN                            NAME   AREA  POP2005 REGION SUBREGION    LON     LAT                       geometry
1   AG   DZ  DZA  12                         Algeria 238174 32854159      2        15  2.632  28.163 MULTIPOLYGON (((2.96361 36....
2   AO   AO  AGO  24                          Angola 124670 16095214      2        17 17.544 -12.296 MULTIPOLYGON (((11.75083 -1...
3   BN   BJ  BEN 204                           Benin  11062  8490301      2        11  2.469  10.541 MULTIPOLYGON (((2.484418 6....
4   CF   CG  COG 178                           Congo  34150  3609851      2        17 15.986  -0.055 MULTIPOLYGON (((12.77905 -4...
5   CG   CD  COD 180 Democratic Republic of the Congo 226705 58740547      2        17 23.654  -2.876 MULTIPOLYGON (((12.95305 -5...
6   BY   BI  BDI 108                         Burundi   2568  7858791      2        14 29.887  -3.356 MULTIPOLYGON (((29.2299 -3....
```

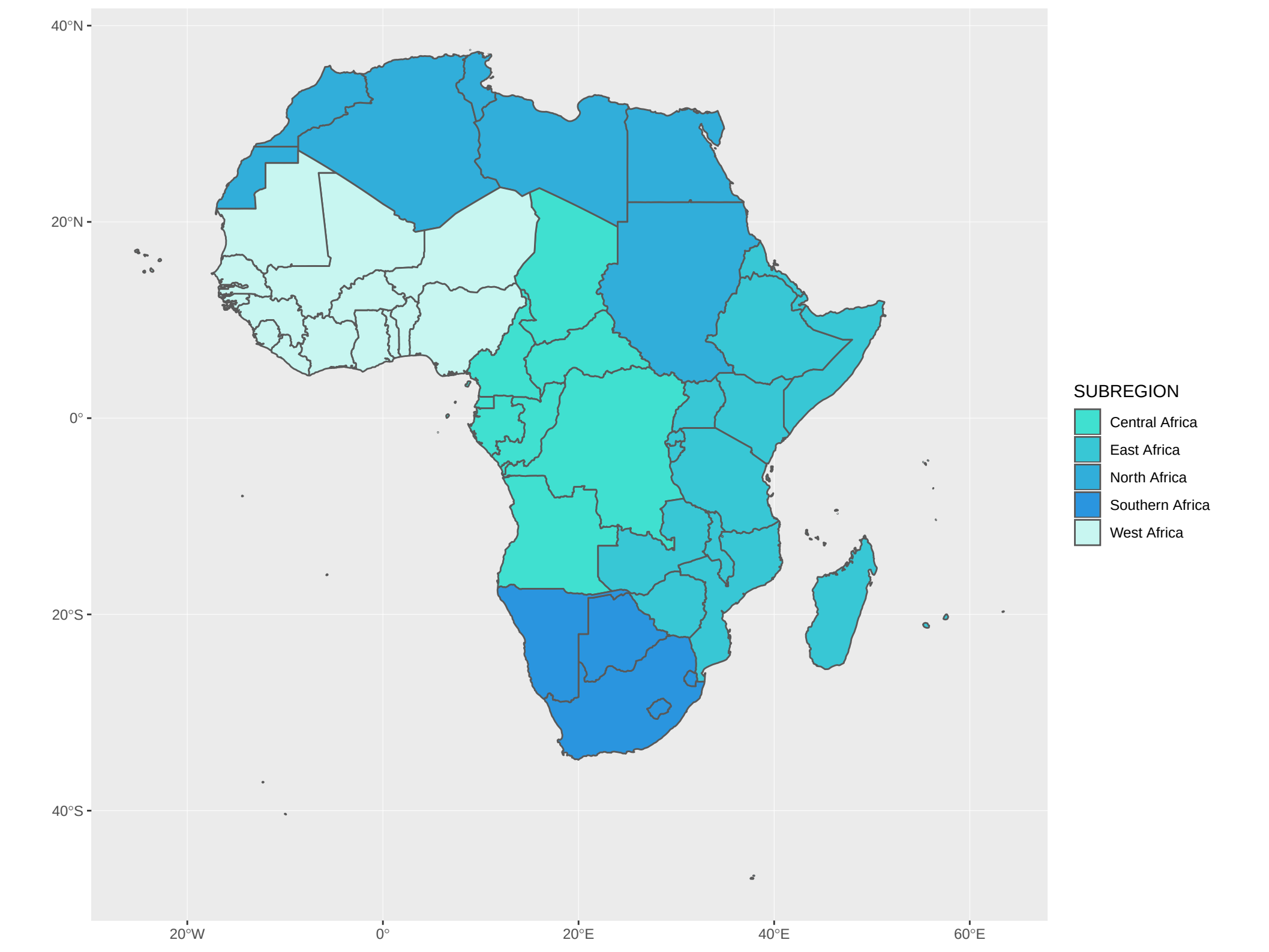
```
# read in sampling info
sampling_info <- (read.csv(paste0(inDir, "/sampling_info.csv")))


# add coordinates derived from map_data
sampling_data <- full_join(sampling_info, sampling_coordinates, by = 'NAME')
```

	NAME	MAG	LON	LAT
1	Uganda	100	32.3860	1.28000
2	South Africa	100	23.1210	-30.55800
3	Ethiopia	120	39.6160	8.62600
4	Nigeria	99	8.1050	9.59400
5	Gambia	113	-15.3860	13.45300
6	Kenya	101	37.8580	0.53000
7	Sierra Leone	85	-11.7920	8.56000
8	Nigeria	109	8.1050	9.59400
9	South Africa	8	23.1210	-30.55800
10	South Africa	8	23.1210	-30.55800
11	Burkina Faso	34	-1.7400	12.27800
12	Mali	50	-3.5240	17.35000
13	Nigeria	49	8.1050	9.59400
14	Ghana	26	-1.2070	7.96000
15	Benin	50	2.4690	10.54100
16	Botswana	48	23.8150	-22.18200
17	Zambia	41	26.3200	-14.61400
18	Cameroon	50	12.2770	5.13300
19	Other	400	102.3446	7.36965
20	Guinea	46	-10.9420	10.43900
21	Cote d'Ivoire	40	-5.5560	7.63200
22	Democratic Republic of the Congo	23	23.6540	-2.87600
23	Uganda	50	32.3860	1.28000

```
# quickly find colours in between

colorRampPalette(colors=c("Turquoise", "Blue"))(10)
```

```
 [1] "#40E0D0" "#38C7D5" "#31AEDA" "#2A95DF" "#237CE4" "#1C63EA" "#154AEF" "#0E31F4" "#0718F9" "#0000FF"
```

```
# plot map_data

ggplot(map_data) + geom_sf(aes(fill = SUBREGION)) +

scale_fill_manual("SUBREGION", values = c("#40E0D0", "#38C7D5",
                                          "#31AEDA", "#2A95DF", "#C8F6F1"))
```

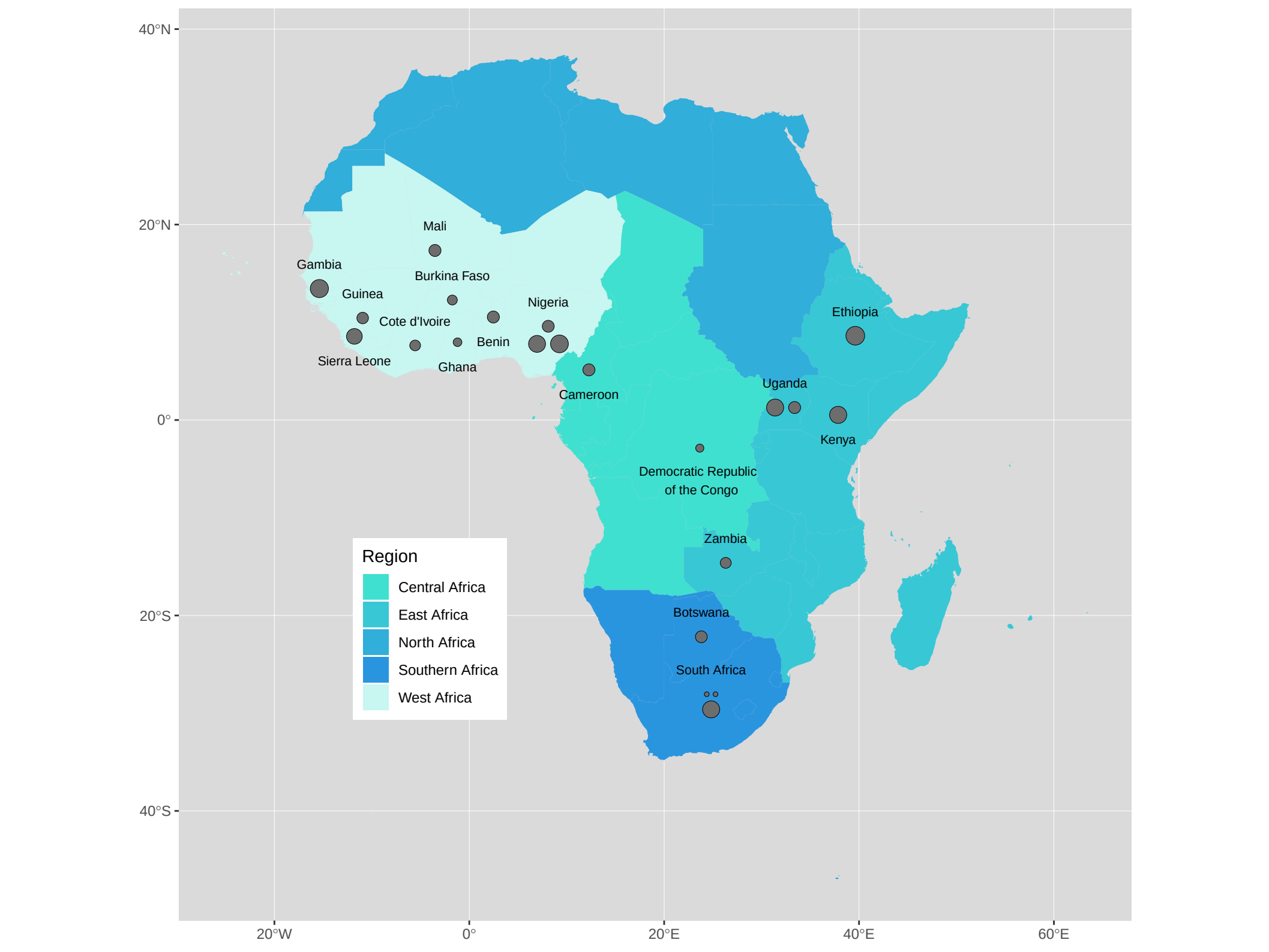

40°N
20°N
0°
20°S
40°S
20°W
0°
20°E
40°E
60°E
SUBREGION
Central Africa
East Africa
North Africa
Southern Africa
West Africa

```r
# add country polygons, with labels above or below country midpoints

p <- ggplot() +

    geom_sf(data = map_data, aes(fill = SUBREGION), color = 'transparent', lwd = 0.5) +

    geom_label(data = map_data, aes(map_data$LON, map_data$LAT, label = map_data$NAME),
                        alpha = 0, size = 2.8, label.size = 0) +

    scale_fill_manual("Region", values = c("#40E0D0", "#38C7D5", "#31AEDA",
                                                        "#2A95DF", "#C8F6F1")) +
```

```
# add country polygons, with labels above or below country midpoints

p <- ggplot() +

    geom_sf(data = map_data, aes(fill = SUBREGION), color = 'transparent', lwd = 0.5) +

    geom_label(data = map_data, aes(map_data$LON, map_data$LAT, label = map_data$NAME),
                        alpha = 0, size = 2.8, label.size = 0) +

    scale_fill_manual("Region", values = c("#40E0D0", "#38C7D5", "#31AEDA",
                                                    "#2A95DF", "#C8F6F1")) +


# specify map variables

q <- p +

theme(legend.justification = c(0.6, -0.1), legend.position = c(0.28, 0.2), axis.title.x = element_blank(),
          axis.title.y = element_blank(), panel.background = element_rect(fill = "#DADADA")
```

```
# add country polygons, with labels above or below country midpoints

p <- ggplot() +

    geom_sf(data = map_data, aes(fill = SUBREGION), color = 'transparent', lwd = 0.5) +

    geom_label(data = map_data, aes(map_data$LON, map_data$LAT, label = map_data$NAME),
                        alpha = 0, size = 2.8, label.size = 0) +

    scale_fill_manual("Region", values = c("#40E0D0", "#38C7D5", "#31AEDA",
                                                    "#2A95DF", "#C8F6F1")) +


# specify map variables

q <- p +

theme(legend.justification = c(0.6, -0.1), legend.position = c(0.28, 0.2), axis.title.x = element_blank(),
          axis.title.y = element_blank(), panel.background = element_rect(fill = "#DADADA")


# add sampling data, scaled by number of participants

r <- q +

    geom_point(data = sampling_data, aes(x = LON, y = LAT), stroke = 0,
                        shape = 21, fill = "#6D6D6D", cex = (sqrt(sampling_data$MAG)/2))
```

```
# add country polygons, with labels above or below country midpoints

p <- ggplot() +

    geom_sf(data = map_data, aes(fill = SUBREGION), color = 'transparent', lwd = 0.5) +

    geom_label(data = map_data, aes(map_data$LON, map_data$LAT, label = map_data$NAME),
                        alpha = 0, size = 2.8, label.size = 0) +

    scale_fill_manual("Region", values = c("#40E0D0", "#38C7D5", "#31AEDA",
                                                    "#2A95DF", "#C8F6F1")) +


# specify map variables

q <- p +

theme(legend.justification = c(0.6, -0.1), legend.position = c(0.28, 0.2), axis.title.x = element_blank(),
          axis.title.y = element_blank(), panel.background = element_rect(fill = "#DADADA")


# add sampling data, scaled by number of participants

r <- q +

    geom_point(data = sampling_data, aes(x = LON, y = LAT), stroke = 0,
                        shape = 21, fill = "#6D6D6D", cex = (sqrt(sampling_data$MAG)/2))


# export figure

r %>% ggsave(filename=paste0(outDir, "/figure.pdf"), width = 28, height = 21, unit = "cm")
```


40°N
20°N
0°
20°S
40°S
20°W
0°
20°E
40°E
60°E
Mali
Gambia
Burkina Faso
Guinea
Nigeria
Cote d'Ivoire
Benin
Sierra Leone
Ghana
Cameroon
Ethiopia
Uganda
Kenya
Democratic Republic
of the Congo
Zambia
Botswana
South Africa
Region
Central Africa
East Africa
North Africa
Southern Africa
West Africa