Proportional visualization of genotypes and phenotypes with rainbow boxes: methods and application to sickle cell disease

Al Hassim Diallo, Gaoussou Camara, Moussa Lo,
Ibrahima Diagne, Jean-Baptiste Lamy
alhassimdiallobe @ gmail.com , jean-baptiste.lamy @ univ-paris13.fr

LANI
Université Gaston Berger de Saint-louis
B.P. 234 Saint-Louis, Sénégal

LIMA
Université Alioune Diop de Bambey
B.P. 30 Bambey, Sénégal

CERPAD
Université Gaston Berger de Saint-louis
B.P. 234 Saint-Louis, Sénégal

LIMICS
Université Paris 13, Sorbonne Universités, INSERM
93017 Bobigny, France

This work is part of the SIMENS project which is supported by the Senegalese High Study and Research Ministry through the Scientific and Technic Research Impetus Funding (FIRST), the African Center of Excellence in Mathematics Computer Science and ICT (CEA-MITIC), and the Pierre Fabre Foundation.

Introduction

- **Screening of genetic disorders**
  - Complex because both phenotype and genotype must be considered
  - **Genotype** : information present in the genome
    - Each patient has two exemplars of each gene (except for chromosome Y)
  - **Phenotype** : observed character (*e.g.* diseased or healthy)
    - Usually resulting from the genotype
- **How to visualize the observed proportion of each phenotype and genotype ?**

The Sickle-cell disease (SCD)

- **Also known as Sickle-cell anemia or Drepanocytosis, is an inherited form of anemia**
  - Characterized by an insufficient number of healthy red blood cells to carry enough oxygen in the body
  - Sickle cell anemia can lead to many complications, including:
    - Acute chest syndrome, Vaso-occlusive crisis, Stroke, Pulmonary hypertension, Organ damage, Blindness, Priapism, Leg ulcers, Gallstones…
- **The need of neonatal screening of SCD**
  - SCD is an inherited disease that affects about 300,000 births worldwide.
  - There are 70 million people affected worldwide, 80% of whom live in sub-Saharan Africa.
  - Both the highest prevalence and highest mortality of sickle cell is in Africa
  - In Senegal, there are no published studies on sickle cell prevalence
  - There is a need
    - 1) for national comprehensive screening to identify patients
    - 2) for developing a holistic care programs to provide therapeutics and education for families and children with the disease

CERPAD

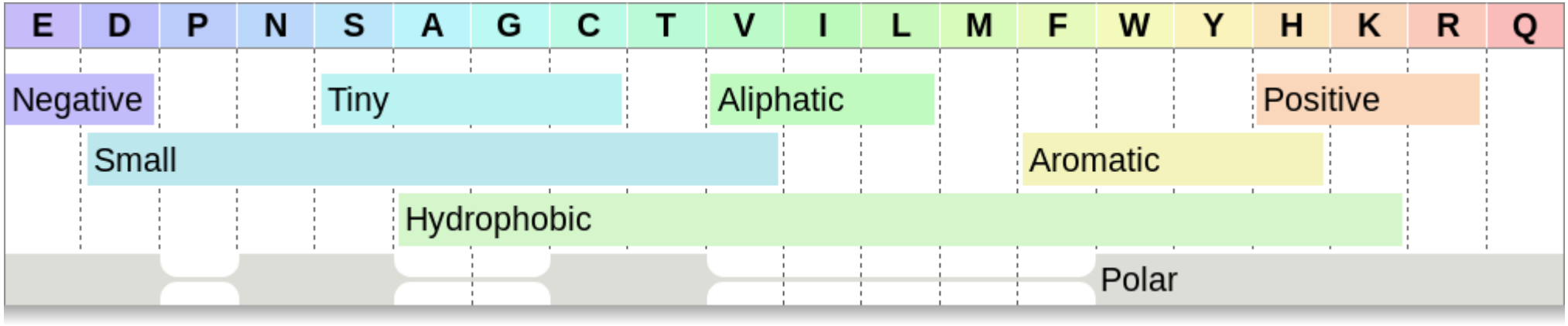
- **Center for Research and Ambulatory Care of the Sickle Cell Disease (CERPAD), Saint-Louis region in Senegal**
  - Funded by the Pierre FABRE Foundation, inaugurated in 2015.
- **Objective: contribute to the fight against sickle cell disease in Senegal**
  - Systematically screen newborns in the maternity wards in the city of Saint-Louis
  - Ensure the follow-up and healthcare of the diseased patient
  - Propose a model for neonatal screening and early care adapted to Senegal's public health system

Genotype and phenotype

- **Genotype : information present in the genome**
  - Each patient has two exemplars of each gene (except for chromosome Y)
- **Phenotype : observed character (e.g. diseased or healthy)**
  - Usually resulting from the genotype
- **Translation as a set visualization problem:**
  - A = { a_1, a_2,… } the set of alleles
  - A genotype is a triplet:
    G = (alleles, proportion, phenotype)
    - G has either 1 allele (both exemplar of the gene are identical)
    - or 2 alleles (two different exemplars)
  - => a set visualization problem in which sets have at most 2 elements

$$\mathcal{A} = \{A, C, S\}$$
$$G_1 = (\{A\}, 30\%, He)$$
$$G_2 = (\{C\}, 12\%, He)$$
$$G_3 = (\{S\}, 11\%, Di)$$
$$G_4 = (\{A, C\}, 7\%, He)$$
$$G_5 = (\{A, S\}, 30\%, Ca)$$
$$G_6 = (\{C, S\}, 10\%, Di)$$

Rainbow boxes

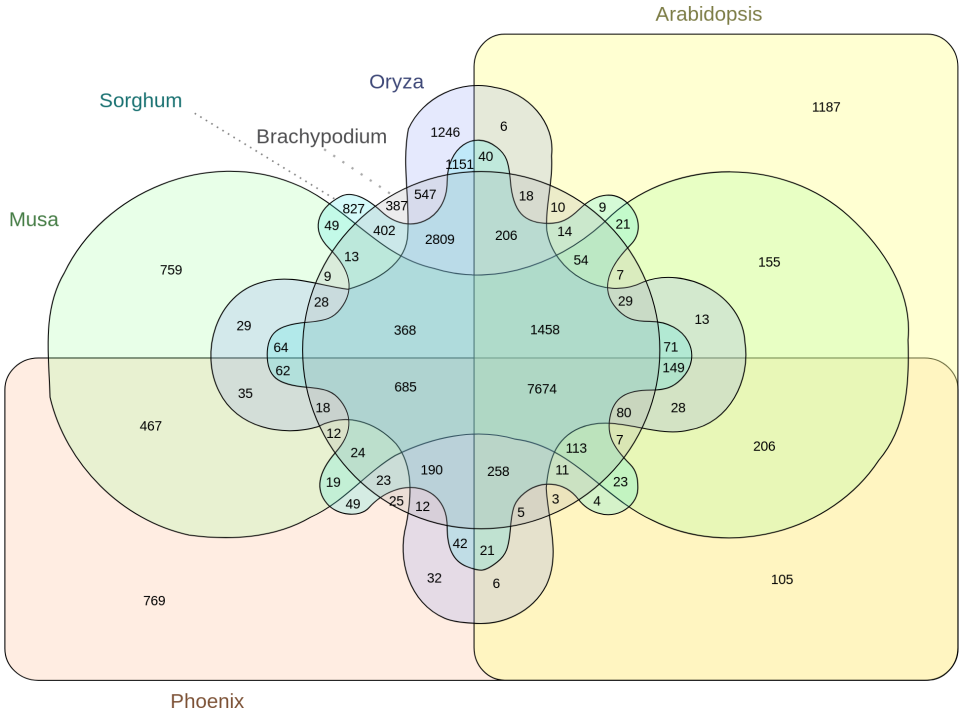


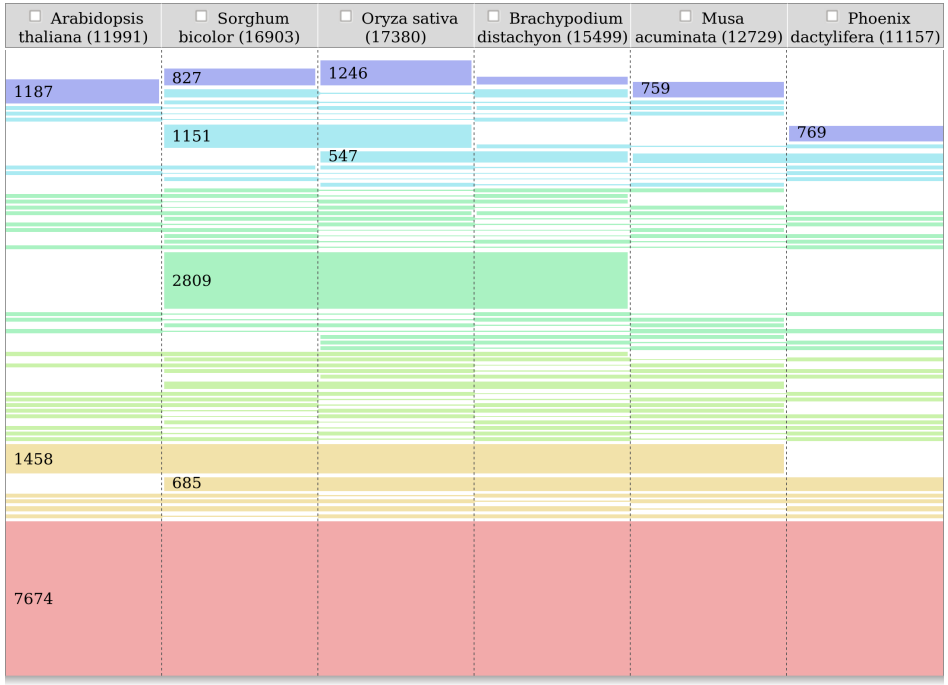
- **Rainbow boxes : a recent technique for set visualization**
  - elements => columns
  - sets => rectangular boxes
  - color => one color per element
  - box color is the mean of its elements color
  - non continguous element in a set => box hole
  - elements are ordered so as to minimize the number of holes
  - box are stacked vertically by size

[Lamy JB et al. Rainbow boxes: a new technique for overlapping set visualization and two applications in the biomedical domain. **Journal of Visual Language and Computing** 2017]

Rainbow boxes

- **Proportional variant : RainBio**
  - Box height encodes cardinality
  - Example in bioinformatics



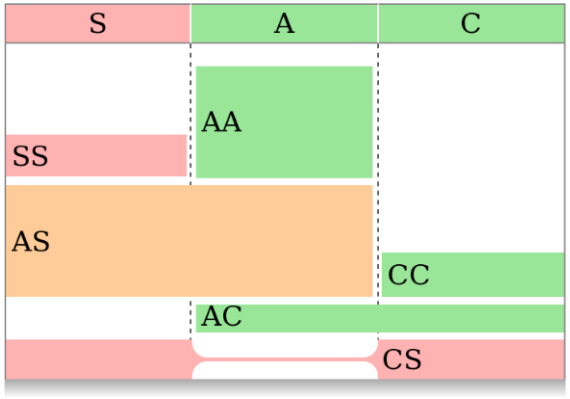


[Lamy JB et al. RainBio: Proportional visualization of large sets in biology. **IEEE Transactions on Visualisation and Computer Graphics** 2019]

Visualization with rainbow boxes

- **Visual encoding**
  - 1 allele => 1 element => 1 column
  - 1 genotype => 1 set => 1 box
  - Genotype proportion => box height
  - Genotype associated phenotype => color
    - Diseased => red, carrier => orange, healthy non carrier => green

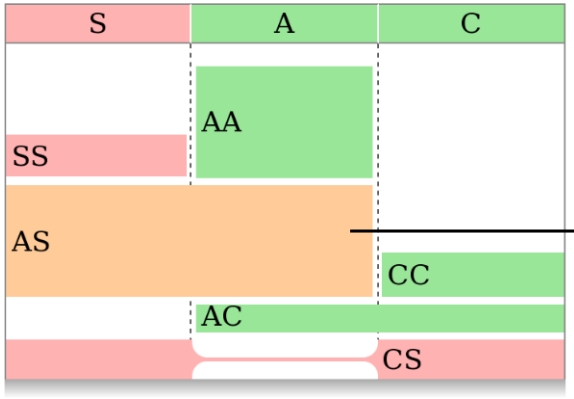
$$\mathcal{A} = \{A, C, S\}$$
$$G_1 = (\{A\}, 30\%, He)$$
$$G_2 = (\{C\}, 12\%, He)$$
$$G_3 = (\{S\}, 11\%, Di)$$
$$G_4 = (\{A, C\}, 7\%, He)$$
$$G_5 = (\{A, S\}, 30\%, Ca)$$
$$G_6 = (\{C, S\}, 10\%, Di)$$



Visualization with rainbow boxes

- **Visual encoding**
  - 1 allele => 1 element => 1 column
  - 1 genotype => 1 set => 1 box
  - Genotype proportion => box height
  - Genotype associated phenotype => color
    - Diseased => red, carrier => orange, healthy non carrier => green

$$\mathcal{A} = \{A, C, S\}$$
$$G_1 = (\{A\}, 30\%, He)$$
$$G_2 = (\{C\}, 12\%, He)$$
$$G_3 = (\{S\}, 11\%, Di)$$
$$G_4 = (\{A, C\}, 7\%, He)$$
$$G_5 = (\{A, S\}, 30\%, Ca)$$
$$G_6 = (\{C, S\}, 10\%, Di)$$

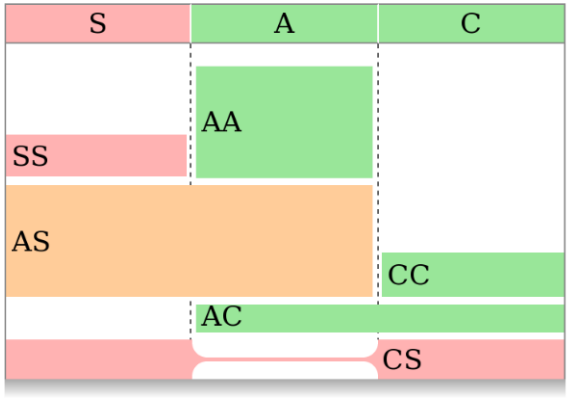


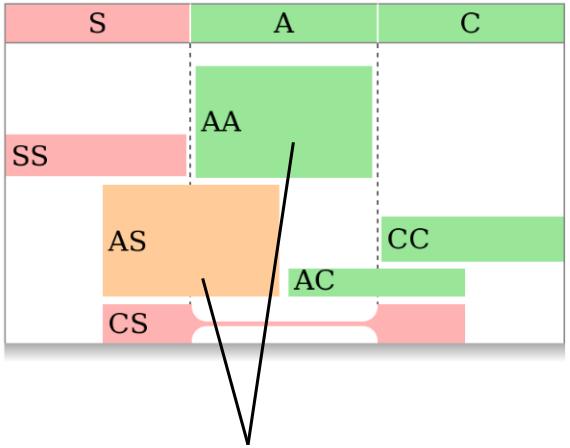
“Are there twice as many AS patients as AA patients ?”

Visualization with rainbow boxes

- **Rainbow boxes improvement for dataset with sets of at most 2 elements**
  - All boxes have the same width
  - Boxes corresponding to sets with 2 elements are in the middle of the 2 columns

$\mathcal{A} = \{A, C, S\}$
$G_1 = (\{A\}, 30\%, He)$
$G_2 = (\{C\}, 12\%, He)$
$G_3 = (\{S\}, 11\%, Di)$
$G_4 = (\{A, C\}, 7\%, He)$
$G_5 = (\{A, S\}, 30\%, Ca)$
$G_6 = (\{C, S\}, 10\%, Di)$



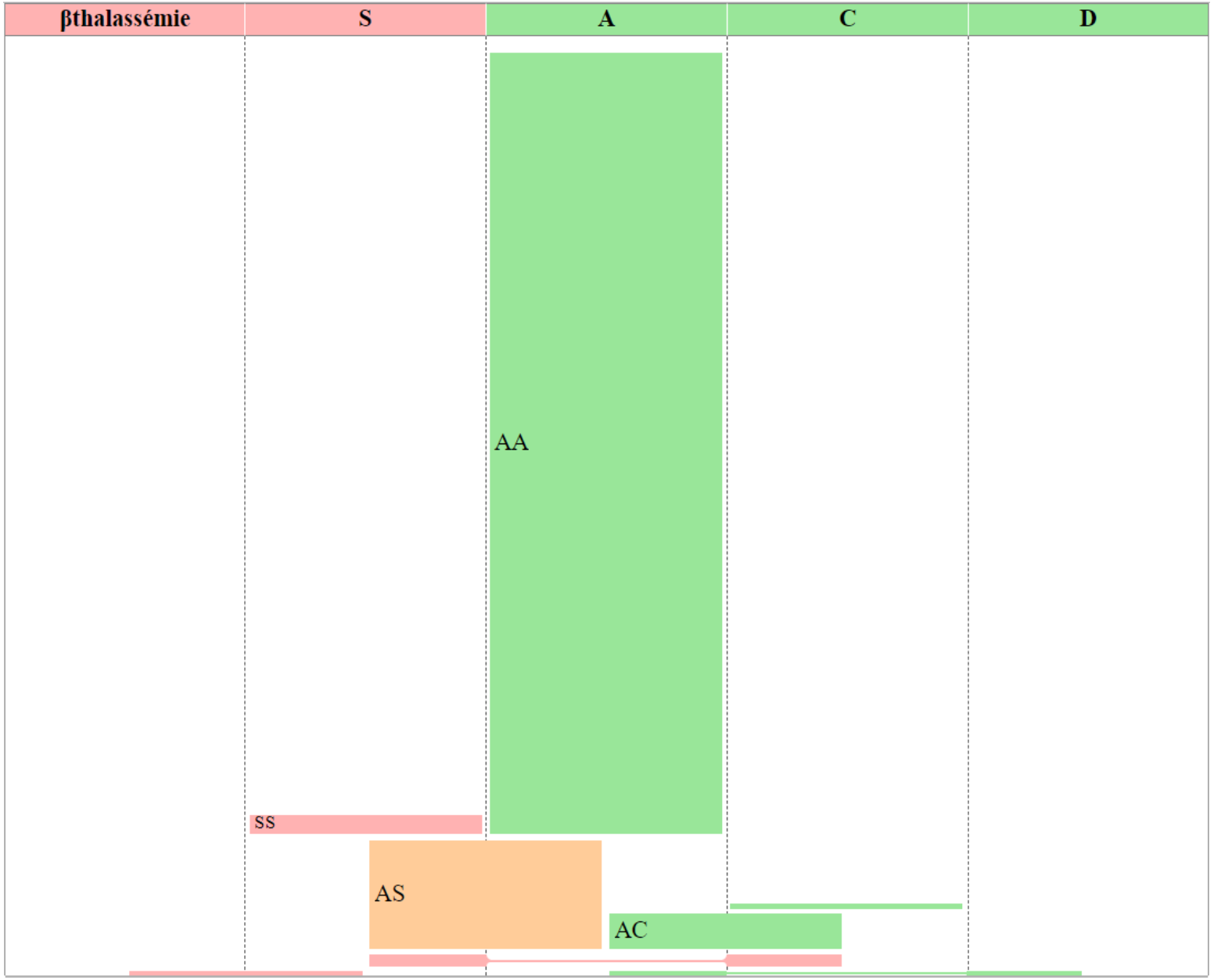


"Same proportion of AS and AA patients"

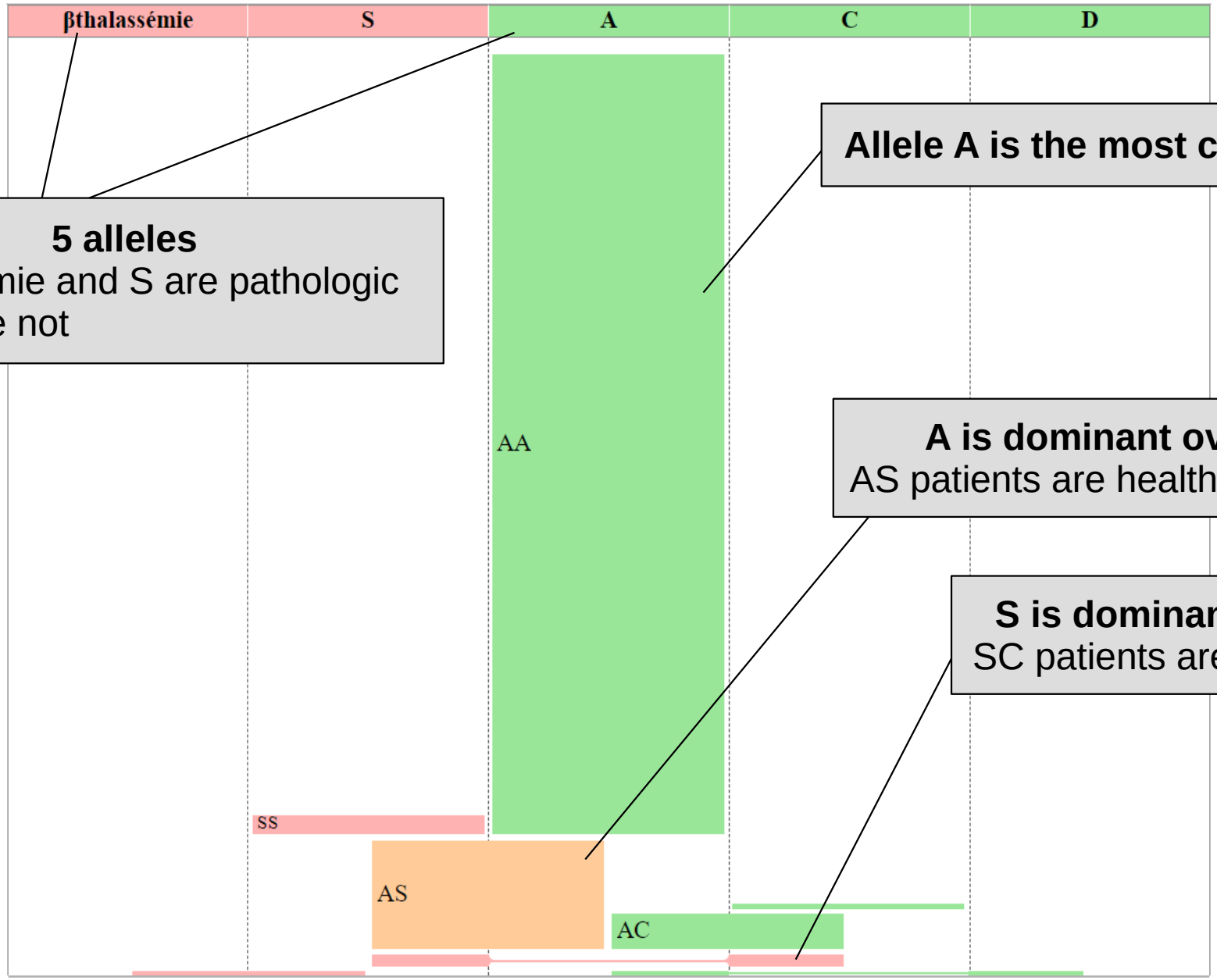
Data collected

- **Sickle cell neonatal screening program in Saint-Louis region of Senegal**
  - From the main maternity ward (CHRSL)
  - It performs about half of deliveries, out of 15 maternity wards
- **5,045 records collected from 25/04/2017 to 26/02/2019**
- **The SIMENS software was used for collecting data**
- **3 main ethnic groups: Wolof, Peulh and Toucouleur**

Application to Sickle-cell disease



Application to Sickle-cell disease



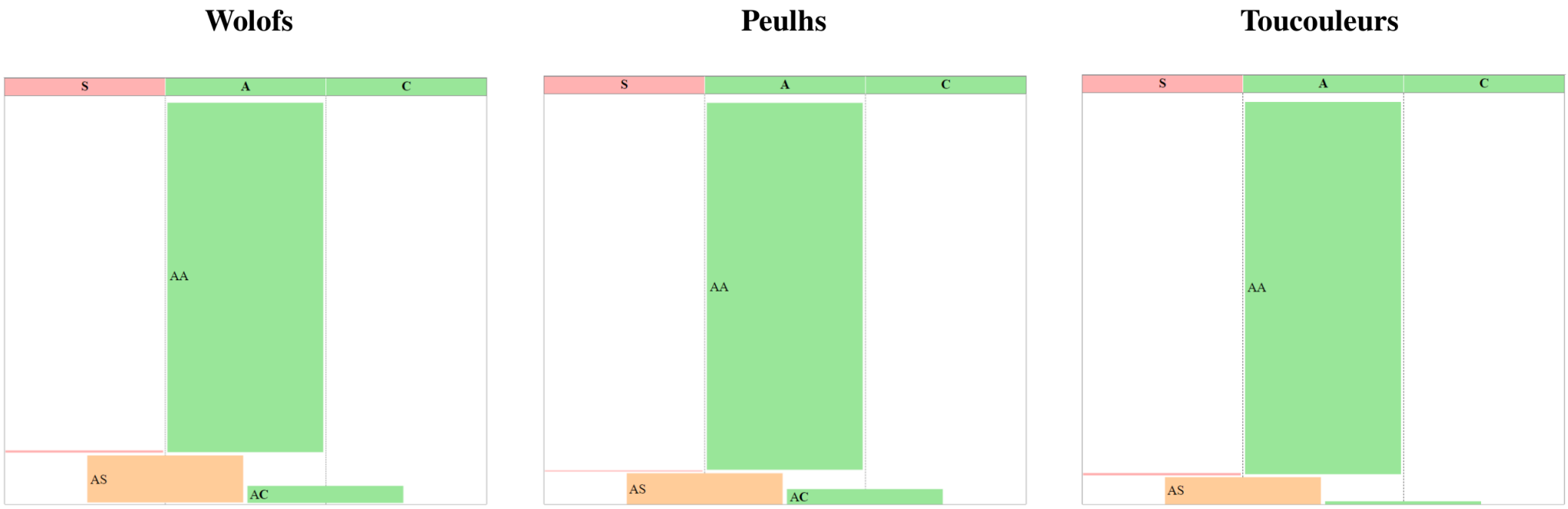
5 alleles
Bthalassemie and S are pathologic
A, C, D are not

Allele A is the most common

A is dominant over S
AS patients are healthy carriers

S is dominant over C
SC patients are diseased

Application to Sickle-cell disease



- **Comparison of the 3 main ethnic groups**
  - Similar overall “big picture”, but:
  - Healthy carriers are more common in Wolofs
  - Allele C is less common in Toucouleurs

Expert opinions

- **2 specialists of sickle cell disease screening at the CERPAD**
  - They found the approach interesting
  - They liked the visualization of alleles, genotypes and phenotypes in a single image
  - Other disorders are related to the same alleles as those found in sickle cell disease
  - => experts suggested the visualization of additional phenotypes

Discussion

- **Set visualization is an original approach for genotype and phenotype**
  - Sets with at most 2 elements
  - In the literature: proportional Venn diagrams, but only approximate
- **For some diseases, the phenotype may not be entirely determined by the genotype**
  - Role of the environment
  - In this case, the box representing a genotype may be split in two parts (a diseased part and a healthy one)
- **For rare diseases, proportional may be very small**
  - => Use a log scale
- **Perspectives:**
  - Integration in SIMENS for the follow-up
  - Application to other genetic disorders
  - Visualization of several phenotypes as suggested by experts
  - Implementation of additional subgroup analyses, *e.g.* sex, countries, geographic areas or maternity wards, socioeconomic groups, time period

References

Lamy JB, Berthelot H, Capron C, Favre M. Rainbow boxes: a new technique for overlapping set visualization and two applications in the biomedical domain. Journal of Visual Language and Computing 2017;43:71-82

Lamy JB, Tsopra R. RainBio: Proportional visualization of large sets in biology. IEEE Transactions on Visualisation and Computer Graphics 2019

Camara G, Diallo AH, Lo M, Tendeng JN, Lo S. A National Medical Information System for Senegal: Architecture and Services. Studies in health technology and informatics 2016

Diallo AH, Camara G, Lamy JB, Lo M, Diagne I, Makalou D, Diop M, Doupa D. Toward an information system for sickle cell neonatal screening in Senegal. Studies in health technology and informatics 2019

Diallo AH, Camara G, Lamy JB, Lo M, Diagne I, Makalou D, Diop M, Doupa D. SIMENS-LIS4SC, a laboratory information system for biological tests of sickle cell screening and healthcare. Studies in health technology and informatics 2019