aviesan
alliance nationale
pour les sciences de la vie et de la santé

Institut Thématique Multi-Organismes
Technologies pour la santé

Etudier la diversité génétique en population générale pour mieux comprendre les maladies

Emmanuelle Génin
UMR1078, Inserm, UBO, EFS, Brest

Inserm
La science pour la santé
From science to health

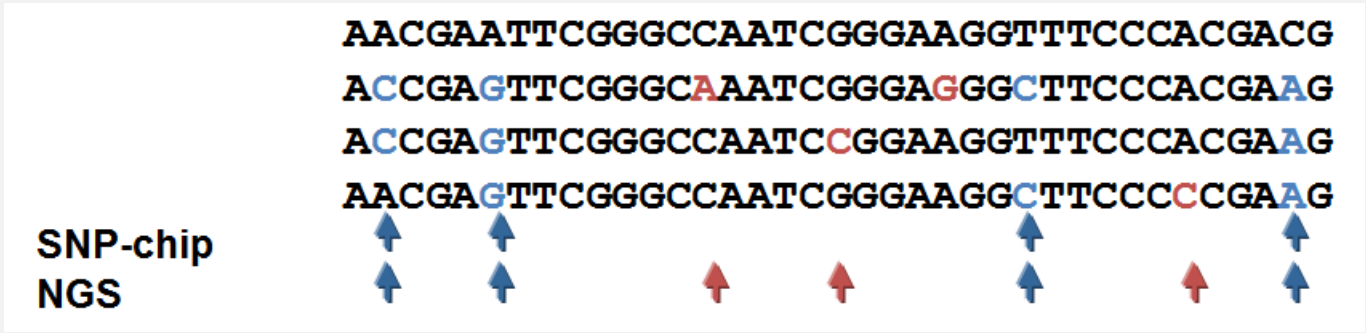
UBO
Université de Bretagne Occidentale

EFS
ÉTABLISSEMENT FRANÇAIS DU SANG
Du donneur aux patients

Le séquençage du génome humain

1977	•W. Gilbert & F. Sanger Première méthode de séquençage	→800 bases en 5 jours
1980	• Prix Nobel de Chimie	
1990	• Composés fluo et automatisation • Projet génome humain	→500 000 bases en 1 jour
2003	• Première séquence du génome	→13 ans et 3 milliards $
2006	• Miniaturisation et nanotechnologies • Séquençage NGS	→1 génome en 1 jour coût ~1000€

Avancées technologiques
Vers une caractérisation plus fine de la variabilité du génome humain

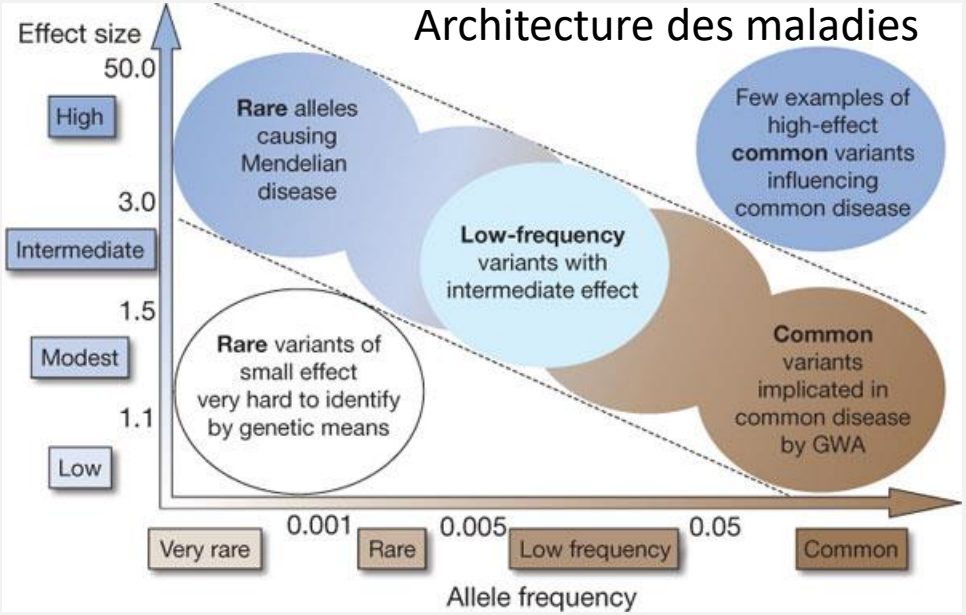


Soin

Médecine Génomique

Recherche

Architecture des maladies



Analyse des données de séquençage

- Pour le **diagnostic de maladies génétiques**
  Analyse individuelle
  - Identifier le variant causal dans le génome de l'individu

- Pour rechercher des **associations dans les maladies complexes**
  Analyse populationnelle
  - Etude cas-témoins
  - Recherche d'un enrichissement en variants rares dans un gène

Filtrer les variants selon leur fréquence

Bases de données publiques

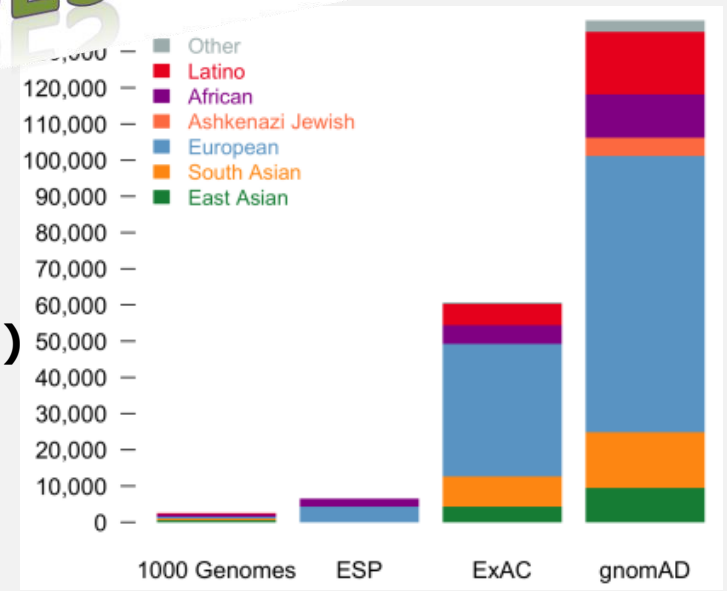
- **1000 Genome Project**
  2504 individus
  26 populations à travers le monde

- **Exome Variant Server**
  NHLBI GO Exome
  6,500
  (

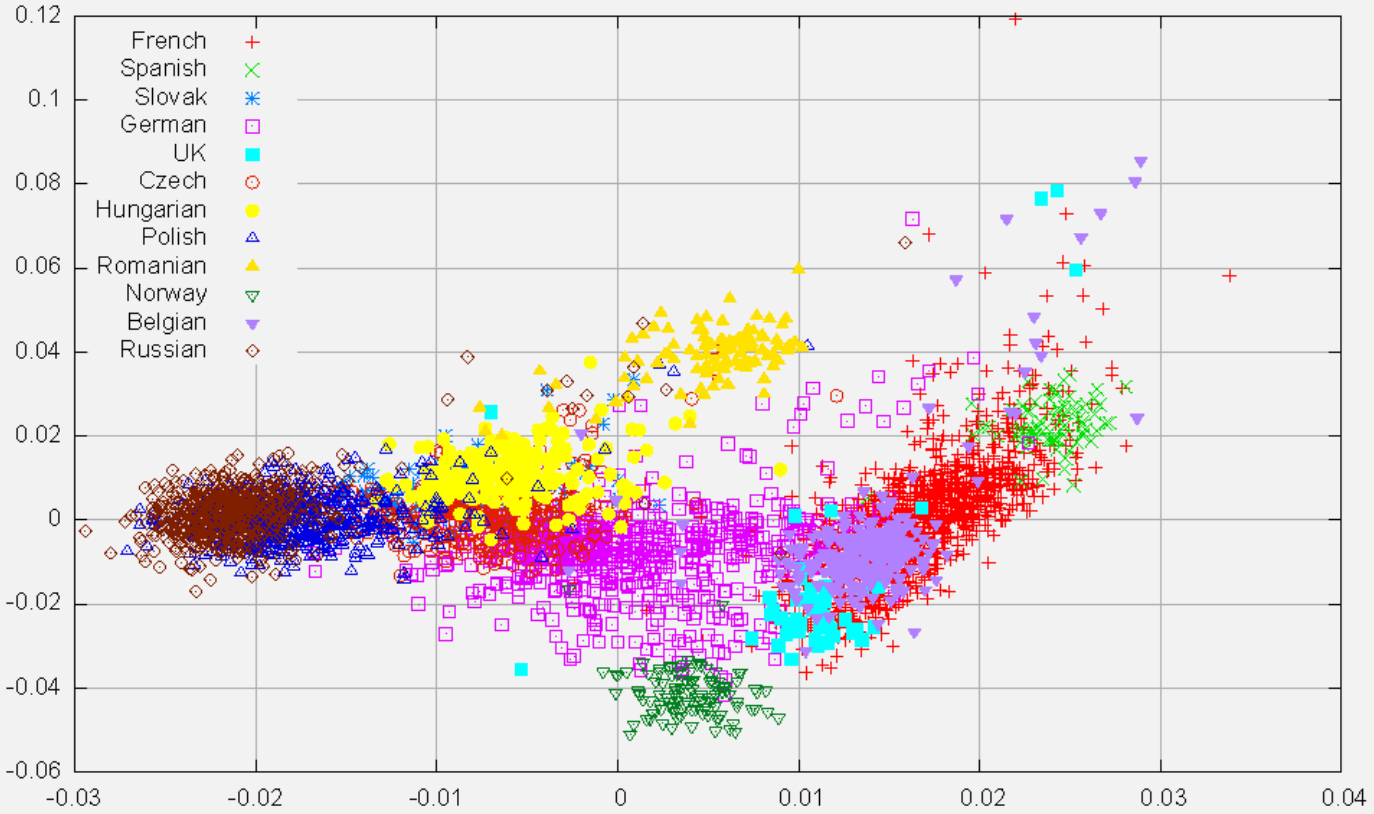
ABSENCE DE DONNÉES FRANÇAISES

- **I**
  M
  (
  ropeens

- **Genome Aggregation Consortium (GnomAD)**
  MacArthur lab Broad Institute
  123,136 exomes et 15,496 génomes
  63,369 Européens



Des differences génétiques selon l'origine géographique

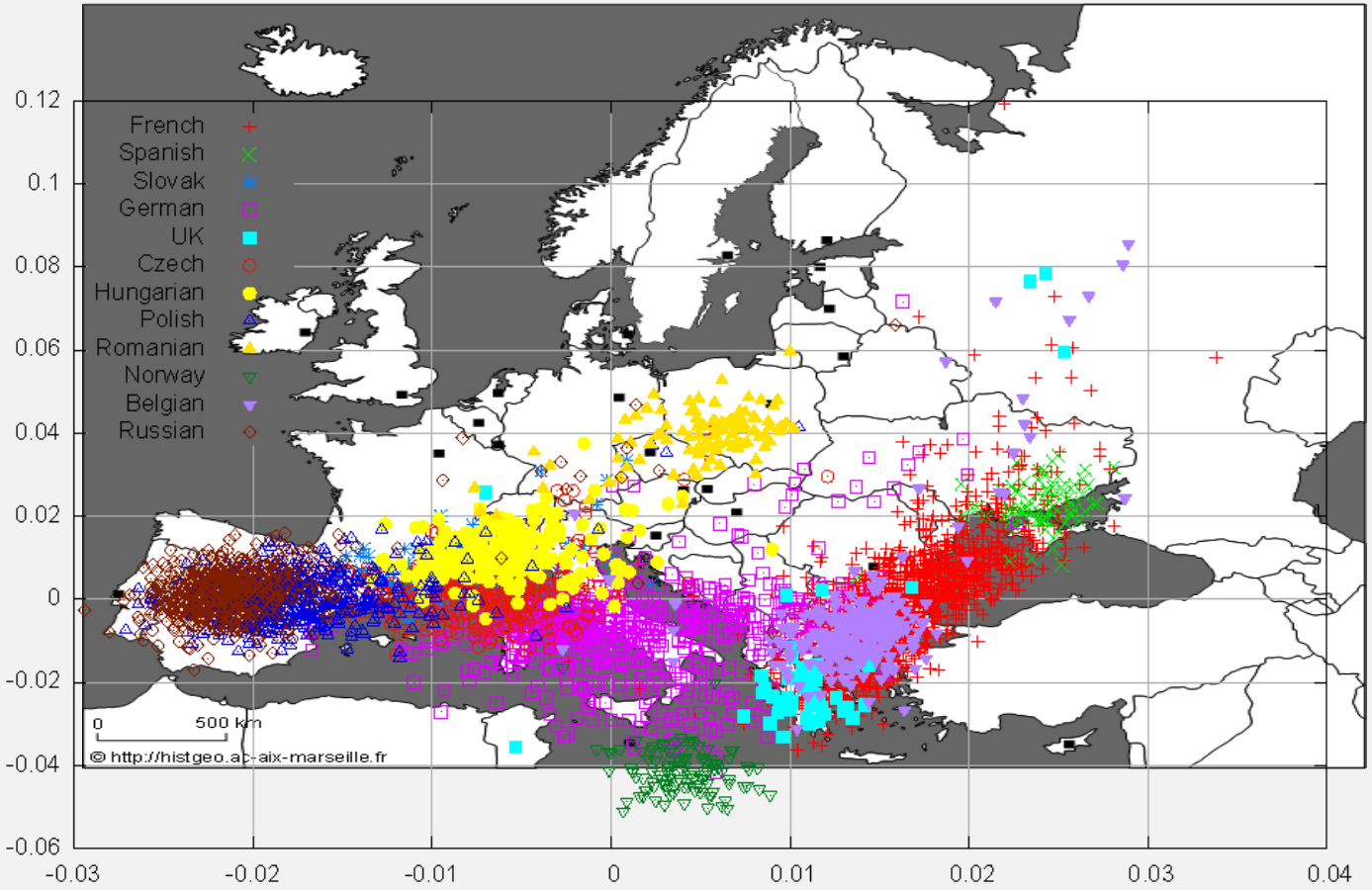


5,811 individus issus de 12 populations européennes

121,242 SNPs

(Heath et al., EJHG, 2008)

Des differences génétiques selon l'origine géographique



5,811 individus issus de 12 populations européennes

121,242 SNPs

(Heath et al., EJHG, 2008)

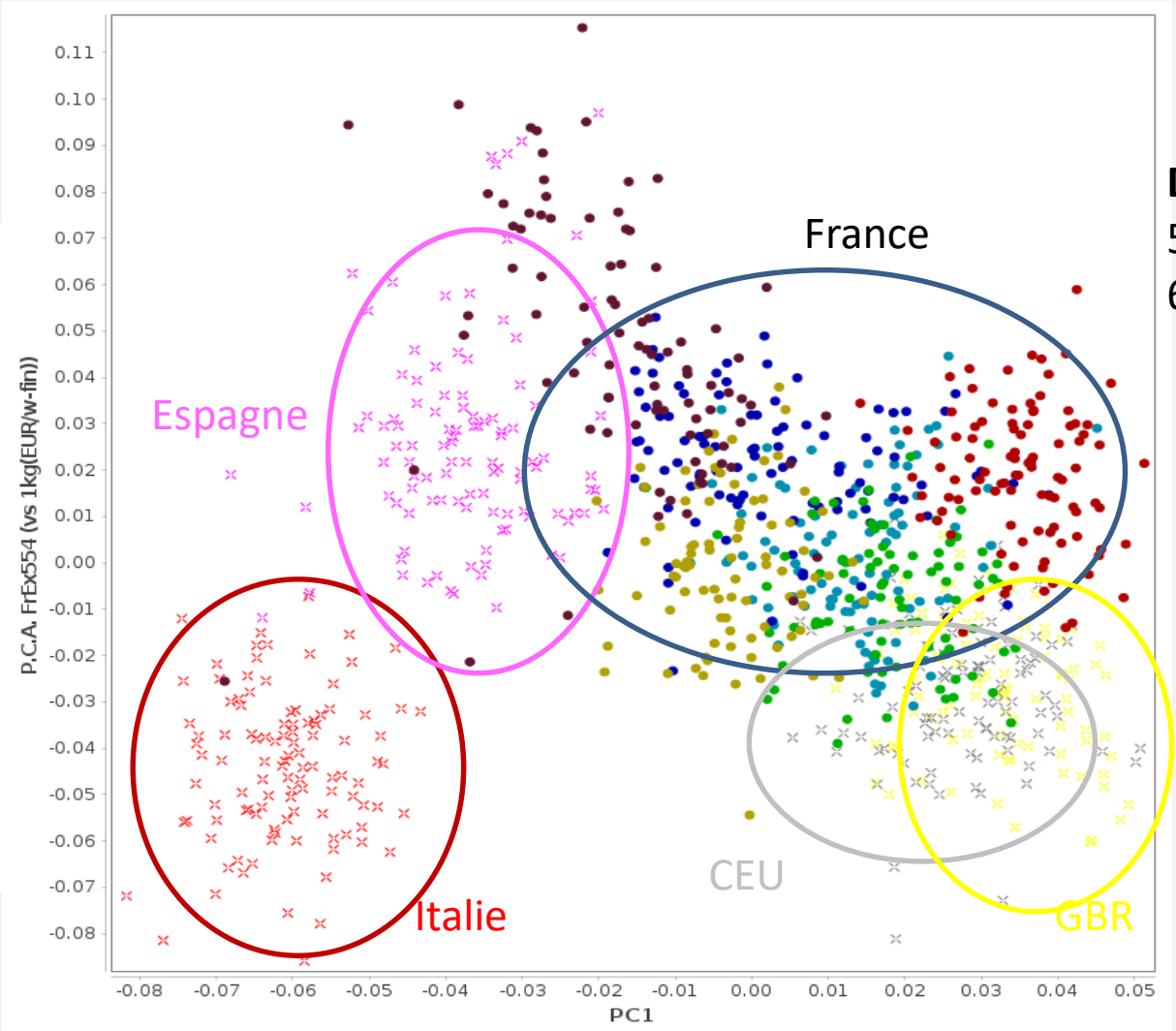
J-F Deleuze
CEA-CNRGH,
Evry

Vers la mise en place de panels de reference

R. Redon
L'institut du thorax
Nantes

- **French Exome Project (FREX)**
  Financement de France Génomique 2013 Call
  574 exomes de 6 régions françaises
- **VaCaRMe (Institut du Thorax, Nantes)**
  Financement de la Région Pays-de-la-Loire
  Population du Grand-Ouest (PREGO)
- **France GenRef Project**
  Financement du Labex GENMED
  ~900 WGS (cohorte GAZEL + PREGO)
- **Projet POPGEN**
  Plan France Médecine Génomique 2025 – MESR
  4 000 WGS (cohorte Constances)

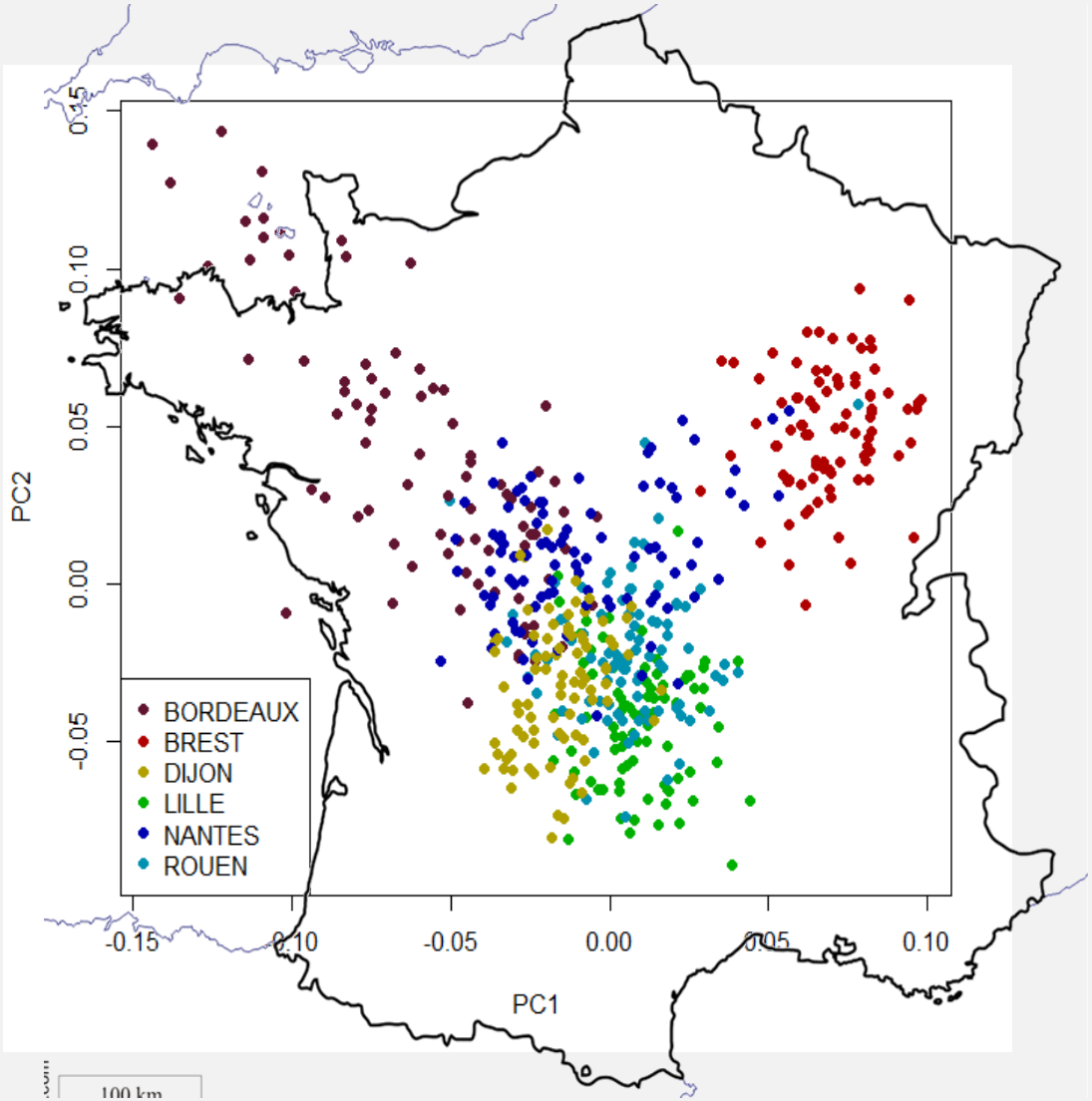
Le panel FREX comparés aux autres panels européens



Données FREX
574 exomes
6 régions

FRANCE GÉNOMIQUE

Variabilité des fréquences alléliques en France

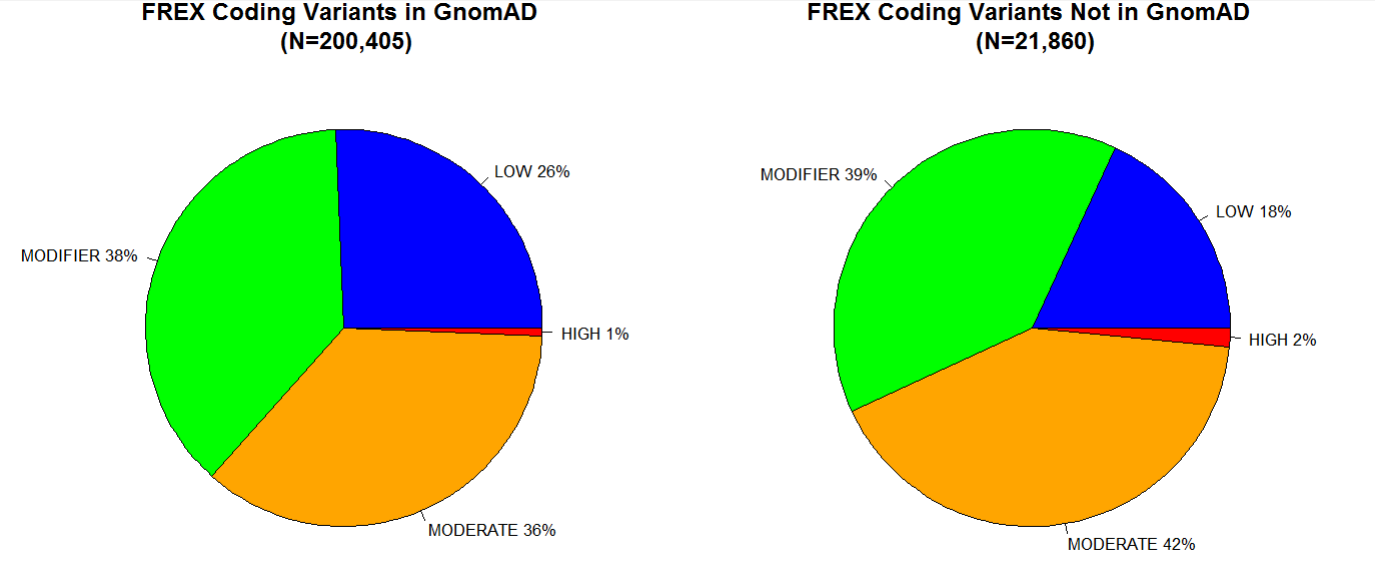


Données FREX
574 exomes
6 régions

FRANCE GÉNOMIQUE

FREX COMPARÉ À EXAC & GNOMAD

- **18.36% des variants codants de FREX (autosomes) sont absents d'ExAC**
  17.97% des SNVs et 36.74% des Indels
- **9.95 % sont absents de GnomAD**
  9.83% des SNVs et 15.45% des Indels
- **Ces SNVs de FREX absents de GnomAD ont des impacts forts**
  Plus souvent annotés par VEP « HIGH » ou « MODERATE »
  43.1% vs 36.6% pour les SNVs présents dans GnomAD
  Plus souvent annotés SIFT « deleterious » et POLYPHEN « prob-damaging »
  25.6% vs 17.6% pour les SNVs présents dans GnomAD



Contribution de Frex a l'interprétation des données NGS de patients

Variants	Total	Absents des Bases de données (%)	Absents des Bases de données et de FREX (%)
Tous les SNVs	100 055	4 775 (4.8%)	2 168 (2.2%)
SNVs Exonic	14 968	168 (1.1%)	81 (0.5%)
SNVs Faux sens	6 450	56 (0.9%)	30 (0.5%)

Valeurs médianes obtenues sur 10 exomes réalisés pour un autre projet sur une autre plateforme de séquençage.

GenMed

FRANCE GÉNOMIQUE

INVESTISSEMENTS D'AVENIR

CONTRIBUTION DE FREX AUX ETUDES D'ASSOCIATION

ORIGINAL ARTICLE

SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease

Molecular Psychiatry (2015), 1–6

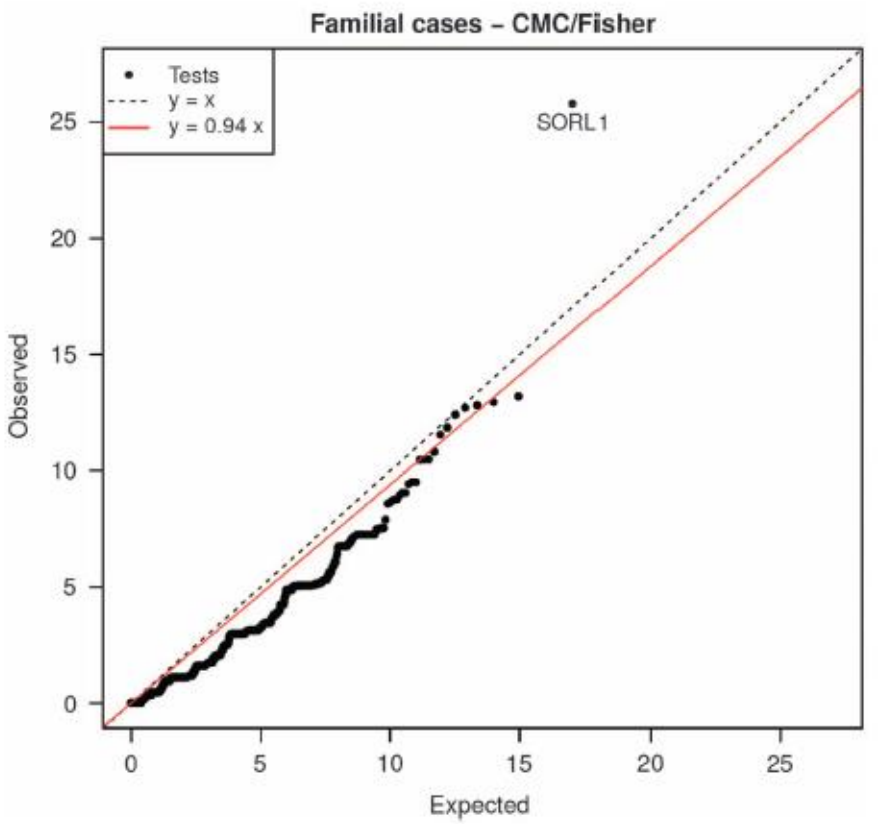


Figure 1. Quantile–quantile plot of gene-level Fisher's *P*-values among 205 early-onset Alzheimer's disease (EOAD) cases with positive family history and 498 controls ($n = 13630$ tests).

CONTRIBUTION DE FREX AUX ETUDES D'ASSOCIATION

European Journal of Human Genetics (2017) 25, 995–1003

www.nature.com/ejhg

ARTICLE

Rare *RNF213* variants in the C-terminal region encompassing the RING-finger domain are associated with moyamoya angiopathy in Caucasians

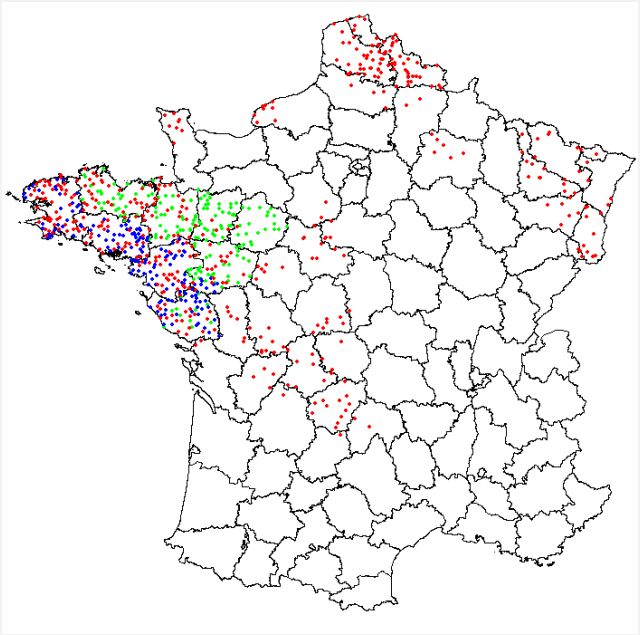
Stéphanie Guey[1], Markus Kraemer[2,8], Dominique Hervé[1,3,8], Thomas Ludwig[4], Manoëlle Kossorotoff[5], Françoise Bergametti[1], Jan Claudius Schwitalla[2], Simone Choi[1], Lucile Broseus[1], Isabelle Callebaut[6], Emmanuelle Genin[4,9] and Elisabeth Tournier-Lasserve*[,1,7,9] the FREX consortium[10]

Moyamoya angiopathy (MMA) is a cerebral angiopathy affecting the terminal part of internal carotid arteries. Its prevalence is 10 times higher in Japan and Korea than in Europe. In East Asian countries, moyamoya is strongly associated to the R4810K variant in the *RNF213* gene that encodes for a protein containing a RING-finger and two AAA+ domains. This variant has never been detected in Caucasian MMA patients, but several rare RNF213 variants have been reported in Caucasian cases. Using a collapsing test based on exome data from 68 European MMA probands and 573 ethnically matched controls, we showed a significant association between rare missense RNF213 variants and MMA in European patients (odds ratio (OR) = 2.24, 95% confidence interval (CI) = (1.19–4.11), $P = 0.01$). Variants specific to cases had higher pathogenicity predictive scores (median of 24.2 in cases versus 9.4 in controls, $P = 0.029$) and preferentially clustered in a C-terminal hotspot encompassing the RING-finger domain of RNF213 ($P < 10^{-3}$). This association was even stronger when restricting the analysis to childhood-onset and familial cases (OR = 4.54, 95% CI = (1.80–11.34), $P = 1.1 \times 10^{-3}$). All clinically affected relatives who were genotyped were carriers. However, the need for additional factors to develop MMA is strongly suggested by the fact that only 25% of mutation carrier relatives were clinically affected.
European Journal of Human Genetics (2017) **25**, 995–1003; doi:10.1038/ejhg.2017.92; published online 21 June 2017

WGS - Projet FranceGenRef
Labex GENMED

Projet	fastq	bam	vcf
FRENCHWGFIN (50 individus)	3,2 To	6 To	206 Go
FRENCHWGPREGO (354 individus)	24 To	40 To	1,3 To
FRENCHWGGAZEL (458 individus)	34 To	49 To	1,6 To
Total	**61,2 To**	**95 To**	**3,1 To**

V. Meyer - CNRGH



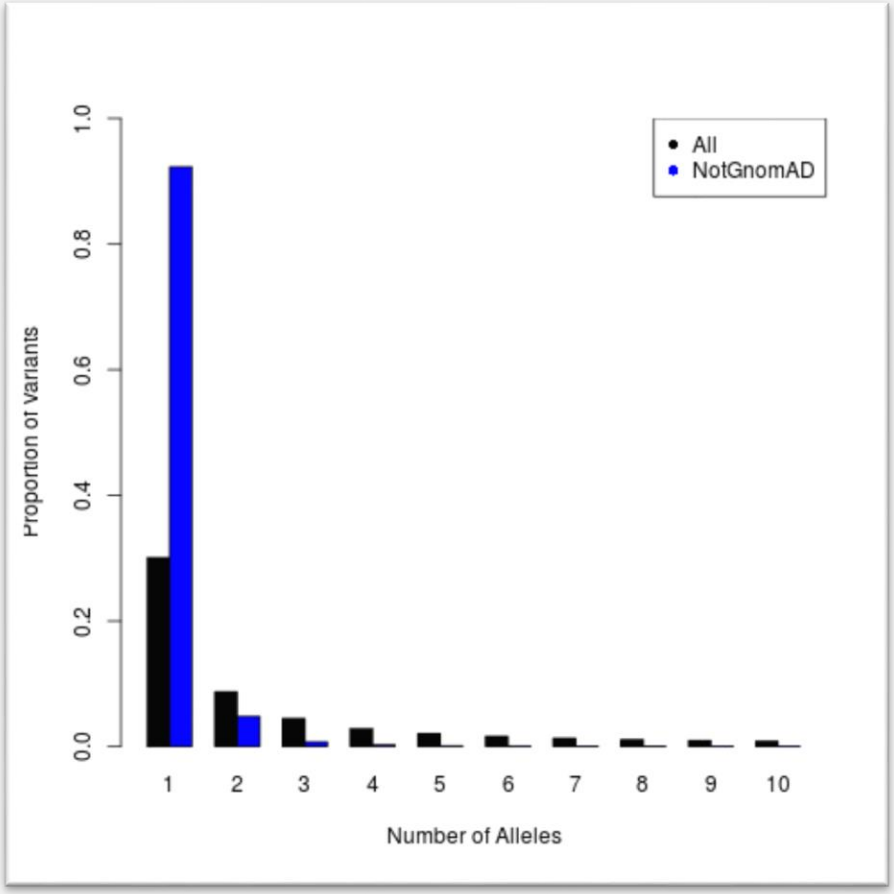
Contrôle qualité: exclusion de 6 individus

- 4 apparentés proches
- 2 ADNs mélangés

	Homme	Femme	Total
PREGO	177	175	352
FINISTERE	24	24	48
GAZEL	384	72	456
TOTAL	**585**	**271**	**856**

M. Karakachoff – L'institut du Thorax

Description des variants
Echantillon total



25,729,497 variants (filtre PASS)
30.1% sont des singletons

6,957,985 absents de GnomAD
92.3% sont des singletons

Description des variants par individu

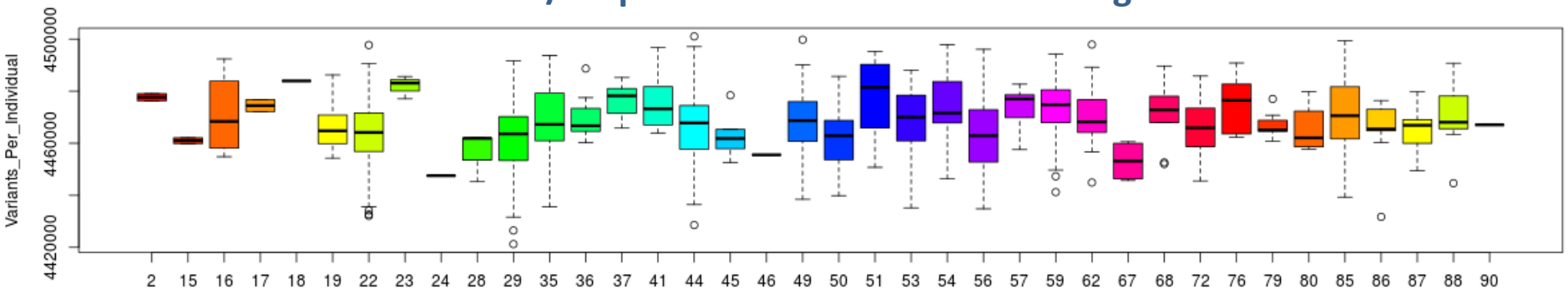
En moyenne: **4,467,190 variants** par individu

dont **14,100 singletons**

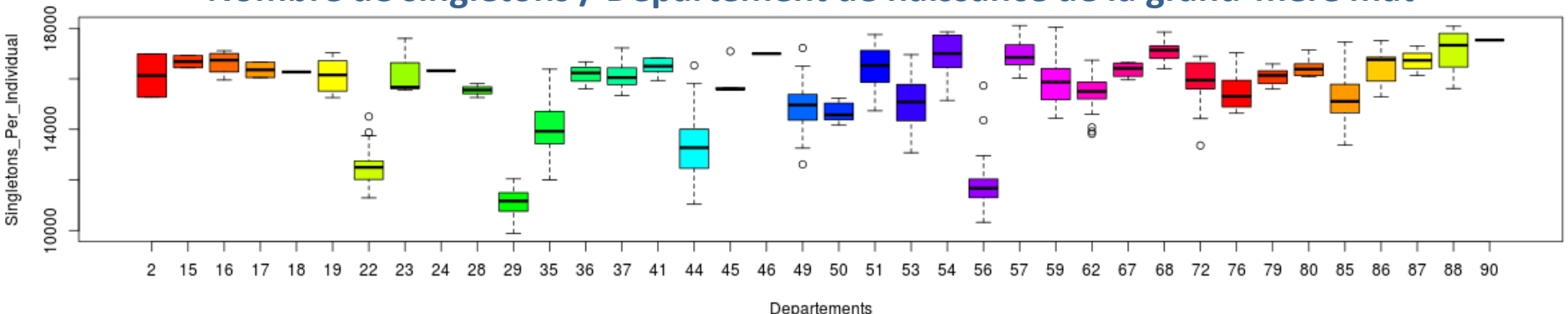
Le nombre de singletons est plus faible en Bretagne:

11,118 pour Finistère vs 14,528 ailleurs ($p<2\ 10^{-16}$)

Nombre de variants / Département de naissance de la grand-mère mat



Nombre de singletons / Département de naissance de la grand-mère mat



Stratification Génétique

Isabel Alves, Joanna Giemza et Christian Dina
l'institut du Thorax, Nantes

- Chromopainter & FineStructure
  1,393,454 SNVs
- 15 clusters avec plus de 10 individus

Partager les données

- Quelles données ?
  - Données agrégées

The FREX browser

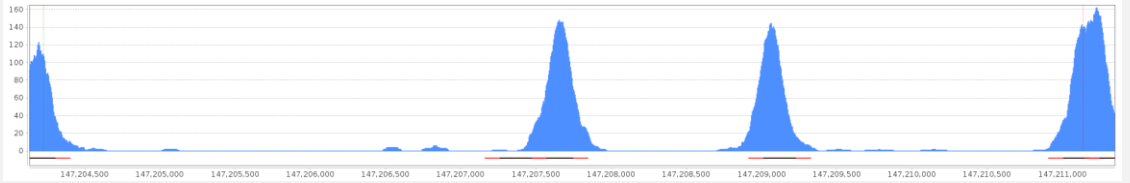
http://med-laennec.univ-brest.fr/FrExAC/

Choose one or more variants

One per line
(ex : "1:1657123", "rs123456789", "ABCA1")

- (•) Show both SNPs and INDELS ?
- ( ) Show only SNPs ?
- ( ) Show only INDELs ?

Ok



Show 10 entries — Search:

Chr	Position	RS ID	Ref	Alt	Protein Csq	Transcript Csq	Annotation	Gene	ExAC AF	ExAC EUR AF	ExAC NFE AF	Type	FrEx AF	FrEx Alt(Homo)/Tota
5	147204093	rs4151639	C	T			Downstream Gene	SPINK1	-	-	-	SNP	0.0252	28(0)/1108
5	147204192	rs11319	G	A		c.*32C>T	3' UTR	SPINK1	0.0769	0.0455	0.0401	SNP	0.0261	29(1)/1108
5	147204266	rs143014431	T	G	p.Lys66Asn	c.198A>C	Missense	SPINK1	0.000148	0.000207	0.000228	SNP	0.000902	1(0)/1108
5	147204290	rs377350168	A	T		c.195-21T>A	Intron	SPINK1	0.000148	0.000154	0.00017	SNP	0.000902	1(0)/1108
5	147204334	rs554919880	G	GAAAA		c.195-69_195-66dupTTTT	Intron	SPINK1	-	-	-	INDEL	0.0126	14(0)/1108
5	147207250	-	C	G		c.194+335G>C	Intron	SPINK1	-	-	-	SNP	0.000922	1(0)/1084
5	147207401	rs114094661	A	T		c.194+184T>A	Intron	SPINK1	-	-	-	SNP	0.0155	17(1)/1094
5	147207616	COSM223032	G	A	p.Pro55Ser	c.163C>T	Missense	SPINK1	0.00446	0.00565	0.006	SNP	0.00542	6(0)/1106
5	147207678	rs17107315	T	C	p.Asn34Ser	c.101A>G	Missense	SPINK1	0.00912	0.0104	0.00973	SNP	0.0126	14(0)/1108
5	147207692	-	C	T		c.88-1G>A	Splice Acceptor	SPINK1	-	-	-	SNP	0.000902	1(0)/1108

Showing 1 to 10 of 19 entries — Previous 1 2 Next

Download results as TSV, CSV, CSV FR (";" instead of ",")

Partager les données

- Quelles données ?
  - Données agrégées
  - Données individuelles

Partager les données

- Quelles données ?
  - Données agrégées
  - Données individuelles

- Comment les partager ?
  - Réaliser des méta-analyses sur les statistiques des tests
  - Echanger des fichiers
  - Mettre en place une plateforme d'analyse sécurisée

→Le projet PRIVGEN

D. Niyitegeka, R. Bellafqira, **G. Coatrieux** – LaTim IMTA Brest
F.Z. Boujdad, M. Sudhölt – LS2N IMT Nantes
T. Ludwig, E. Génin – UMR1078 Brest

Le projet PRIVGEN

Poster de Reda Bellafqira

Poster de Thomas Ludwig

PRIVGEN
Privacy-preserving sharing and processing of genetic data

CominLabs

LaTIM Inserm UMR 1101
LS2N CNRS UMR 6004
Inserm UMR 1078

in collaboration with

Labex Genmed

Partners

Context

- **Cloud Computing and data outsourcing** - A successful paradigm to flexibility store, share and process large amount of data while minimizing costs
- **Security needs of outsourced applications and data are worsened**
  - Owners loss the control on their data and applications (***confidentiality***, ***integrity***, ***availability***?)
  - Service provider may in turn transmit data to third-party service providers (***traceability***, ***intellectual/scientific ownership protection***?)
  - Storage by the service providers of data issued from different sources (***privacy***?)
- **Sharing of outsourced genetic data and applications – more than an experimental framework**
  - Needs for international sharing of genetic data for better human genome decryption to improve diagnosis ...
  - Data highly personal, covering a large security spectrum needs (*privacy, data reliability – integrity + authenticity -, scientific ownership ...*)
  - Distributed applications
  - Different initiatives (e.g. beacons) with identified security weaknesses ...

Objectives

- **Respond to actual security solutions limitations**
  - Cloud applications impose satisfying many security properties at once → Needs to make interacting different security mechanisms
  - Cloud applications are distributed computations executed on behalf of multiple stakeholders
- **Two research axis**
  - Composition of security and privacy mechanisms applied to compositions of complex computations
  - New multipurpose security mechanisms able to satisfy several security objectives at once.

Challenge 1 - Mechanisms for a continuous digital content protection

- **Objective:** Merging different security mechanisms into one configurable digital content protection tool for multipurpose security purposes.
- **Contributions :** Provide continuous data protection with joint security mechanisms configurable by a composition language.

Processing of encrypted genetic data

- **Objective:** Allow two or more research teams to perform genetic association studies while preserving data confidentiality and privacy.
- **Contributions :** Homomorphic encryption based genetic association study using secure χ^2 test.

Controlling the integrity of encrypted genetic data

- **Objective:** Allow the cloud to control the integrity of homomorphically encrypted outsourced data.
- **Contributions :** A dynamic joint homomorphic encrytion-watermarking scheme able to detect and identify altered data under user data update constraints.

Challenge 2 - Composition of security and privacy-protection mechanisms

- **Objective:** Provide a development approach for privacy-preserving distributed genetic applications
- **Contributions :** A composition theory for security and privacy properties - Programming support.

Sharing architecture

- **Contributions :** A multi-cloud based architecture with a trusted party for data processing. Geneticists' data storage is delegated to the Clouds which are independent and non communicating for privacy reasons.

Composition theory

- **Algebraic laws:** extend the theory for security mechanisms combination (watermarking, encryption, fragmentation) with classical queries for correct security query formulation.

Implementation: an abstract implementation in Idris shows the exchange workflow and security operations to perform a GWAS-like analysis in the suggested architecture.

Challenge 3 - Distributed processing of genetic data

- **Objective** - a platform for: i) sharing relevant genomic information while maintaining privacy; ii) supporting the distributed execution of applications over shared genetic data.

People

- G. Coatrieux, Pr., LaTIM Inserm UMR 1101, IMT Atlantique
- M. Südholt, Pr., Ascola, LS2N CNRS UMR 6004, IMT Atlantique
- E. Genin, Dr., Inserm UMR 1078
- J-F. Deleuze, Dr., CNG
- D. Niyitegeka, Ph.D Student, LaTIM Inserm UMR 1101, IMT Atlantique
- F. Bouddad, Ph.D Student, Ascola, LS2N CNRS UMR 6004, IMT Atlantique
- R. Bellafqira, IR, LaTIM Inserm UMR 1101, IMT Atlantique
- T. Ludwig, IR, Inserm UMR 1078

Publications

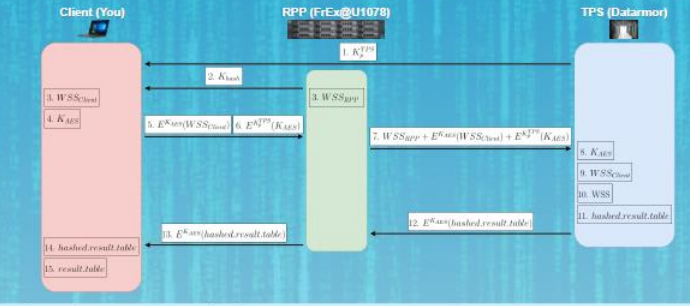
- FZ Bouddad, M Südholt, *Constructive Privacy for Shared Genetic Data - CLOSER 2018*-8th International Conference on Cloud Computing and Services Science, 2018.
- J. Franco-Contreras, G. Coatrieux, *Protection of Relational Databases by Means of Watermarking: Recent Advances and Challenges*. Advances in Security in Computing and Communications, Intechopen, pp. 101-123, 2017.

ANR

PrivAS: a tool to perform Privacy-Preserving Association Studies

Thomas E. Ludwig[1,2], Reda Bellafqira[3], David Niyitegeka[3], Daniel Salas[4], Isabelle Perseil[4], Gouenou Coatrieux[3] and Emmanuelle Génin[1]

PrivAS is a tool to perform Genome-Wide Association studies (GWAS) using the Weighted-Sum Statistic (WSS) algorithm in a Privacy-Preserving environment. The underlying scenario takes into account three interacting parties: (1) a Client, e.g. a genomic research unit, wanting to measure the association between an observed phenotype and regions of the genome; (2) a Reference Panel Provider (RPP) possessing genetic data for a Reference Panel, e.g. a priori healthy individuals of a carefully selected ancestry and (3) a Third-Party Server (TPS) with large computational capacities. Our tool and its underlying implementation preserve both state-of-the-art performances and Privacy for all parties. Indeed, through a series of hashing and encryption mechanisms, we can assure that no genetic data from neither the Client nor the RPP are visible by the other parties involved. Furthermore, only the Client is able to view a decrypted version of the WSS results.

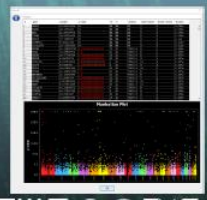


In our implementation of the secure WSS, three parties are involved:

1. the Client that possesses data for individuals presenting the studied phenotype
2. the Reference Panel Repository (RPP) that has data for unaffected individuals
3. the Third-Party Server (TPS) that will do the actual computation.

In order to allow these parties to work together without compromising the privacy of the data, encryption and hashing mechanisms will be implemented. The TPS will execute the *WSS* algorithm data where the variant its gene's name have been hashed, using the *SHA256* algorithm initialized with a key K_{hash} shared by the Client and the RPP but unknown to the TPS. As the Client doesn't have direct access to the TPS, its data will transit through the RPP server. Since the RPP knows K_{hash}, it is able to intercept the Client's data. So, these data are encrypted using the *AES* algorithm with a key K_{AES} generated by the Client. As the TPS needs to be able to decipher the Client's data, the Client sends K_{AES} to the TPS via the RPP, protecting the key from RPP by using an *RSA* encryption. The Client uses the public *RSA* key from the TPS K_1^{TPS} and encrypts K_{AES} with it. Later the TPS uses its secret *RSA* key K_2^{TPS} to decrypt the message. Once all computations are done, the TPS sends the results (that contain hashed gene names and their estimated p_{value}) to the Client via the RPP. The results are encrypted using the *AES* key K_{AES} from the Client. Finally, the Client decrypts the results and unhashes the gene names.

1. Client gets RSA K_1^{TPS} from TPS
2. Client gets the session's unique SHA256 hash key K_{hash} from RPP
3. Client and RPP use K_{hash} to hash variants and gene names, producing WSS_{Client} and WSS_{RPP}. Client builds hash dictionary
4. Client generates a unique AES key K_{AES}
5. Client uses K_{AES} to encrypt WSS_{Client} and sends $E^{K_{AES}}(WSS_{Client})$ to RPP
6. Client uses K_1^{TPS} to encrypt K_{AES} and sends $E^{K_1^{TPS}}(K_{AES})$ to RPP
7. RPP sends WSS_{RPP}, $E^{K_{AES}}(WSS_{Client})$ and $E^{K_1^{TPS}}(K_{AES})$ to TPS
8. TPS uses RSA K_2^{TPS} to retrieve K_{AES}
9. TPS uses K_{AES} to retrieve WSS_{Client}
10. TPS performs WSS association tests for each $hash^{K_{hash}}(gene)$
11. TPS produces a $hashed.result.table$, listing each $hash^{K_{hash}}(gene)$ to its WSS p_{value}
12. TPS uses K_{AES} to encrypt $hashed.result.table$ and sends $E^{K_{AES}}(hashed.result.table)$ to RPP
13. RPP sends $E^{K_{AES}}(hashed.result.table)$ to Client
14. Client uses K_{AES} to retrieve $hashed.result.table$
15. Client uses hash dictionary on each $hash^{K_{hash}}(gene)$ to get $result.table$



1 Inserm, Univ Brest, EFS, UMR 1078 GGB, F-29200 Brest
2 CHRU Brest, F-29200 Brest
3 UnivB INSERM UMR 1101 LaTIM, IMT Atlantique, Brest
4 INSERM DSI - CSI, F-75013, Paris, France

http://lysine.univ-brest.fr/privas

This work was supported by the labex COMINLABS as part of the PrivGen project

Et la suite…

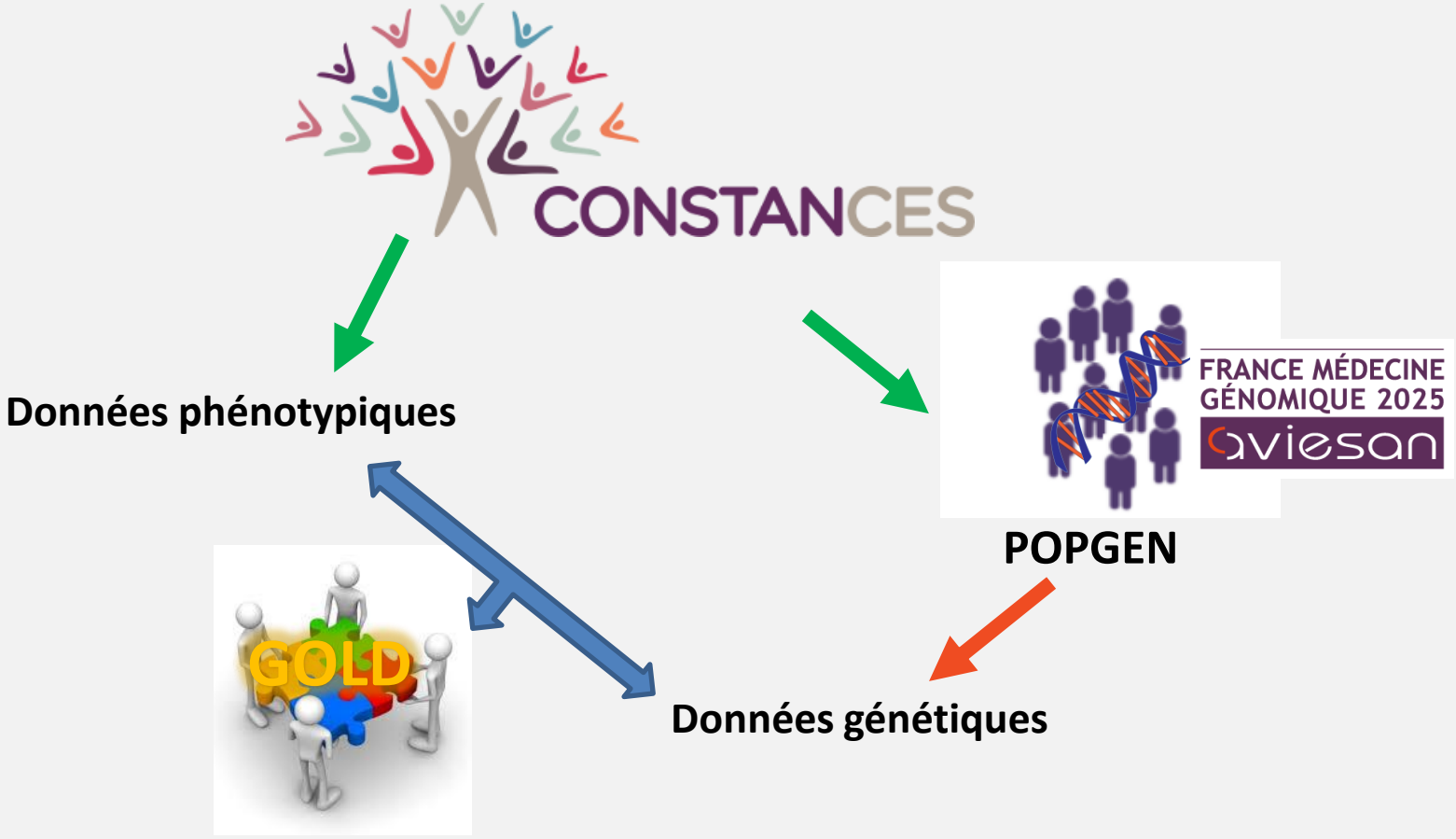
Projet Pilote POPGEN
Variabilité de la population générale

- **Augmenter le panel de référence** en s'appuyant sur la cohorte Constances
  - Sélection de volontaires originaires des différentes régions françaises
  - Couplage des données génétiques avec les données de suivi longitudinal
  - Intégration dans la base de métadonnées (CAD) du plan FMG

Le programme transversal inserm «Variabilité Génomique»

Projet GOLD

Exploiter les données de séquences du projet POPGEN
pour **étudier l'impact des variants génétiques sur la santé**



REMERCIEMENTS

- **UMR 1078, Inserm, UBO, EFS, Brest**
  - Thomas Ludwig
  - Gaëlle Marenne
  - Ozvan Bocher
  - Aude Saint-Pierre
  - Karen Rouault
  - Claude Férec
- **CNRGH/CEPH**
  - Jean-François Deleuze
  - Anne Boland
  - Vincent Meyer
  - Delphine Bacq
  - Hélène Blanché
- **Institut du Thorax, Nantes**
  - Richard Redon
  - Christian Dina
  - Pierre Lindenbaum
  - Matilde Karakachoff
  - Joanna Giemza
  - Isabel Alves
  - Stéphanie Chatel
  - Hervé Le Marec
- **Labex GENMED**
- **Marcel Goldberg, Marie Zins et l'équipe GAZEL/CONSTANCES**
- **Les volontaires GAZEL**