Ancestral genomic submicroscopic inversions of human genome and their relation with multifactorial human diseases

Armand Gutiérrez Arumi

TESI DOCTORAL UPF / 2015

DIRECTOR DE LA TESI

Dr. Luis Alberto Pérez Jurado (Departamento Genomica UPF/IMIM/CIBERER)

Department of Experimental and Health Sciences UPF/IMIM/CIBERER

Agraïments

Quan comences un projecte així tens la sort de conèixer gent molt diferent. I precisament, aquest fet, és un dels elements més enriquidors quan fas una tesi. Si a més es fa amb un centre com aquest a on hi ha gent tant potent encara més. La diversitat certament suma a tots els nivells.

Vaig començar la meva etapa al PRBB com a programador del grup de la Núria López (Biomedical Genomics Group / GRIB) amb un equip de persones excel·lent. A nivell de enginyers, un luxe haber pogut treballar amb en Jordi Deu, Xavier Rafael i Christian Pérez. I un luxe també amb tota la gent amb un knowledge més profund en biologia com l'Abel González, l'Alba, la Sofia, la Gunes i en Mihi, que de fet ells també van col·laborar en despertar més l'interés cap a la bioinformàtica. De fet en aquest grup les barreres disciplinàries es dilueixen i es treballa com un tot per resoldre els projectes. Els informàtics escolten i aprenen dels biòlegs, i els biòlegs dels informàtics (que no faltin scripts). Merci gent!

Quan es va acabar el projecte amb el BG Group, vaig tenir la sort d'anar a parar al grup d'en Luis Pérez Jurado de l'àrea de genètica. Es tracten amb pacients i dades reals, i moltes d'aquestes estan al congelador de -80ªC. La major part de companyes (totes) són de «poiata», i el fet de poder veure que s'amaga darrera de l'explotació de dades genòmiques, que llencem amb tanta soltura i que després cal validar. Realment hi ha una gran feina i dedicació. Que si l'RNA es degrada, que si se'm contaminen les mostres, que si el buffer, que els «FISHos» no surten, que no és una PCR és una esllapissada, que si no punxo bé la neurona, ... i com superen totes aquestes adversitats amb tenacitat, i amb el somriure de premi final, quan aconsegueixen l'objectiu. Començant per la zona a on t'estic, moltes gràcies especialment a la Roser Coromines que ha estat tota una mentora, i a la resta de magnífiques companyes del laboratori Marta, Judit, Tina, Aida, Raquel, Debora, Clara, Ivon i Marivi, que m'han estat ajudant fins el final. A més de resoldre dubtes de biologia, i gaudir d'un munt de pastissos (i coca, i roscos ;)) tots plegats, hem pogut disfrutar junts dels partidillos a volley (i a sobre guanyar un munt de trofeus, poca broma!). I sense els vostres ánims en els últims

metres, no sé si hagués arribat … :) [I a l'*Al Pacino* d'un "*domingo cualquiera*" també, per la part que li toca :)]

En Jairo, Lluis, Xavi, i en Benja de Qgenomics per compartir somriures, problemes técnics i dubtes bioinformàtics.

També a la quarta planta, gràcies pels bons moments (sobretot a la crítica hora de dinar) a Pau, José Luis, Maria, Inés,Cinta, David, i Albert. Gent ha set un fart de riure poder estar amb vosaltres.

An en Juan Ramon González i a tota la gent del CREAL per les BRGE sessions. Especialment a l'Alejandro Cáceres, amb qui he tingut la sort de treballar amb el tema de les inversions, pels bons moments que hem passat i la dedicació per explicar-me de manera fàcil alguns conceptes matemàtics que veia ininteligibles, com els que ha utilitzat per detectar i genotipar les inversions. La comprensió que tinc de l'estadística que ara tinc no seria la mateixa sense ell.

Agraeixo a la fundació IMIM per haver dotat aquesta tesi d'un ajut per a la seva impressió.

Finalment, agraïr al meu director de tesi, en Luis Pérez Jurado per haver confiat en mi, i donar-me la oporunitat d'emprendre aquest projecte personal, i tot el suport que he rebut per part seva. En Luis viu la ciència. És increïble l'accesibilitat que dóna, si no está amb algú altre o enganxat el teléfon, sempre es presenta totalmen disponible per parlar.

A la meva família que sempre m'han apoïat i donat el seu suport amb tot el que he fet. Amb això si que m'ha tocat la loteria.

Finalment vull agrair a la la principal i gran «culpable» de ficarme en tot aquest «barullo», la meva estimada companya Miriam Ortiz. Quan m'explicava el que feia (i fa) al laboratori hem deixava (i deixa) al·lucinat (i com no davant de plantes que produeixen insulina, o que tenen incorporades gens de proteines florescents de medusa? Give me this Hy5! xD).

A tots vosaltres, moltes gràcies!,

Summary

Significant advances were performed in mapping and characterizing different types of structural variation in the human genome such as point mutations, insertions, deletions, etc. Nevertheless, there is still an important genetic component in multifactorial diseases that it is still unknown. Inversions are chromosome mutations in which a given loci change its orientation respect to a reference and present a balanced genetic material (neutral copy number). Our group developed tools in order to infer some variants of ancestral submicroscopic inversions (0.5 Mb – 5Mb) using computational methods with SNP arrays, and documented that inversions correlates with genetic expression.

Recent studies shows that, some inversions, regarding to be neutral in copy-number, could affect to the expression of certain genes. Therefore, It is thought that submicroscopic inversions could have a contributive effect in certain diseases, there are not properly studied.

Our target was try to determine the contribution of those inversions, with the phenotype of four complex disease models: Autism Spectrum Disorders (ASD), Schizophrenia Spectrum Disorders (SSD), Rheumatoid Arthritis (RA) and Ulcerative Colitis (UC), in available datasets (~27,000 samples). Our results have determined that:

- The inversions 17q21.31 and 8p23 are related to ASD and SSD risk. Concretely, 17q21.31 are associated with multiplex ASD families (OR=1.20 p-value 9.52e-05). Meanwhile the inverted allel (17q), gives SSD risk (OR=0.86 p-value 0.007).
- We better characterize 15q24 inversion found three putative inversion-induced haplotypes (NI, Ia and Ib). Homozygous for NI correlated with over-expression of *MAN2C1* over many brain areas, where Ib homozygosity was highly under-expressed.
- RA and UC, we found runs of homozygosity regions (RA OR=9.32 95%CI [2.66-63.45]) and 3p24.3 (UC OR=7.77 95% CI [2.08-54.56]). Additionally, we found haplotypes that could be inversions that are already associate to them, but we are pending to further validations.

Resum

S'han realitzat grans avenços en caracterització dels diferents tipus de variació en el genoma humà, com ara mutacions puntuals, insercions, delecions, etc., però encara existeix un component genètic de les malalties, determinada per l'heretabilitat, que és desconeguda.

Les Inversions són mutacions cromosòmiques en què un fragment de cromosoma canvia la seva orientació respecte a la de referència i són difícils de genotipar per ser neutres en nombre de còpies. El nostre grup ha desenvolupat una eina per inferir algunes variants d'inversions submicroscòpiques ancestrals (0.5MB - 5MB) per mètodes computacionals utilitzant microarrays d'SNPs, i ha documentat que es correlacionen amb expressió gènica.

El nostre objectiu ha estat intentar determinar la contribució de les inversions cromosòmiques al fenotip en quatre models de malalties complexa: autisme, esquizofrènia, artritis reumàtica i colitis ulcerosa. Realitzat estudis de cas control i de segregació en trios, els nostres resultats han determinat que:

- Les inversions 17q21.31 i 8p23 s'associen a risc d'autisme i esquizofrènia. Concretament, la 17q21.31, s'associa especialment amb aquells individus provinents de famílies multiplex (OR=1.20 p-value 9.52e-05). Per altra banda, l'al·lel contrari (17q) confereix risc d'esquizofrènia (OR=0.86 p-value 0.007).
- Hem caracteritzat tres haplotips relacionat amb la inversió de la regió 15q24 (NI, Ia i Ib) que segueixen un patró d'herència Mendeliana amb una alta frequència en Europeus (NI=20%, Ia=50%, Ib=30%). L'homozigositat per l'al·lel no invertit correlaciona amb la sobre-expressió de MAN2C1 en varies àrees del cervell (mentres que l'al·lel Ib està sota expresat).
- Pel que fa a l'artritis reumàtica (RA) i la colitis ulcerosa (UC), hem trobat unes regions d'homozigositat 4q32 (RA OR=9.32 95%CI [2.66-63.45]) i 3p24.3 (UC OR=7.77 95% CI [2.08-54.56]) . Addicionalment, també hem trobat uns haplotips que podien ser inversions que també hi estarien associats però estan pendents de validacions addicionals.

Preface

"Nothing in biology makes sense except in the light of evolution.". Theodosius Dobzhansky

Origin of the disease

If we carefully contemplate our body it is natural to end up with a fascination of what we are seeing. Our hand, for example it is composed by several living tissues, such as epithelial tissue, that protects the muscle and bone tissues, with blood vessels that gently feeds those millions of cells that compose it. Our eyes, for example, with their complexity, with the cornea having the exact amount of curvature, the lens adjusting distance and iris the brightness, so that the optimal quantity of light focuses exactly on the surface of the retina, and the nerves that conducts this information into the brain, to finally perceive an amazing range of shapes and colors. Trillions of cells are composing our organism and works in, apparently, perfect harmony.

This sense of perfection suddenly goes away when we notice that maybe we are not so perfect when we realize that we experience aging processes, and we are susceptible to aging diseases or we question ourselves if some organs should be designed better. For example, we should argue that maybe our knee design was not the best design as possible, compared with for example the knees of the birds or posteriors legs of a Cheetah (*Acinonyx jubatus*), that have an articulation reversed compared as humans. Some engineering studies notice that the human should experience some improvement in taking profit of the force generated and less articulation tension that if the knees were inverted. Other example is the same as pointed by Richard Dawkins in his book Blind Watchmaker (1986) in which he argues that the design of the human eye maybe it is not the best design possible. The reason is that human eye has an initial constraint in which the photoreceptors are pointed away from the light and the wires (that also attenuates part of the light) are concentrated in the middle of the eye forming the optic nerve and creating a blind point. The squid eye on the other hand, as equal to vertebrates, seems to have a more reasonable design according an engineer that is pointing directly to the light. Besides that, the eye nerve contains an inverted image and

travels to opposite parts of the brain, but that is another story. Moreover, it seems that this design of human eye, makes more susceptible for detached retinas, while the squid eye is free of such problems.

Why this process of evolution, that produces these impressive organs, does not avoid that we can suffer a heart attack or a stroke? Why permit this kind of steady deterioration and eventually death? Where is the origin of the diseases? Why our own organism could not avoid an auto-provoked sort of cancer?

Precisely this is the main focus by an interesting article of Randolph M. Nesse and George C. Williams (1988), that focuses on the evolution and origins of disease. They propose several hypotheses that could explain this questions, and finally coined the term Darwinian medicine (or evolutionary medicine). That is applying modern evolutionary theory to understanding health and disease. It looks for understand why the body is designed in a way that makes us all vulnerable to problems like cancer, depression, autoimmune disease, etc., and offer a wider context to conduct research. They propose two main considerations to understand the disease. The first dimension could be explained by evolved defenses. Some discomforting conditions such as fever, cough, pain, vomiting, inflammation and anxiety are in fact, a product of evolved effects of our immune system. For example, a cough is very useful to clear foreign matter from the lungs, and avoid to death by pneumonia. The capacity of pain is also beneficial to sense those things that could go wrong in our organism and let us to take some kind of action. Fever leads to higher body temperature and it facilitate the destruction of pathogens.

Other explanation that those authors comment is in the inherent constraints of the design and selection of several organisms. For example the design of our eyes promotes detached retinas, whereas in squids, that follows other kind of evolution, have nicely avoid. Other example of bizarre designs should be that our respiratory and food passages intersect because in an early lungfish ancestor that filters water and eats zooplankton. A nice example in which show that natural selection cannot start from scratch, and because of that, we have to live with the possible that some food will clog the opening to our lungs.

The other dimension could be explained by the existing conflicts with other organisms. Natural selection is unable to provide us a perfect protection against all

pathogens, because they also are evolving, and it seems that they are doing much faster than humans do. For example, usually bacteria have rapid rates of reproduction, and therefore, much opportunity for mutations in one day, as humanity in a millennium.

With the previous two considerations, and taking into account that the adaptations have to deal with several constraints, they suggest that adaptation makes compromises and tradeoffs, and the diseases could be understood in this way. For example, much suffering is unnecessary but inevitable to detect in a proper response time some kinds of problems. The cost of a false alarm, for example a strong reaction such vomiting in absence of a real thread, is temporary unpleasant. But it is largely compensated if a death is avoided (for example, in expulsing toxic food). As pointed by this two authors, the evolution of a certain organism is working selecting those scenarios in which the cost involves minor inconveniences, and avoiding those that are fatal, such as death before the reproduction of the individual. Moreover, compromise is inherent in every adaptation, for example, if our specie evolved in having several times much thickness in arm bones, we certainly could have almost unbreakable arms, but at the same way would create a great demand on calcium.

This whole set of phenotypic innovations were produced at genetic level in form of a different types of alterations (deletions, insertions, and duplications among others), and therefore were genetically heritable. A nice review of those advantageous variants could be found on Radke, D. W., & Lee, C. (2015), and includes, among many others, an increase of bone mineral (deletion on APC), increase in salivary amylase production (duplication of AMY1), reduced susceptibility of HIV-1 (duplication of CCL3L1), among many others. Unfortunately, most of the alterations end up with a disease, the other side of the coin (Mc Carthy et al. (2008)).

Certainly the evolution does not care about the suffering of the organism or future consequences. Only cares about the reproduction and selection and mutation of some genes. From this point of view, in one study, reprinted in 2001 by Science, of George C. William, suggested that genes that cause aging and eventual death, could be selected that gave some kind of an advantage in youth, when the force of

selection is stronger. He explains that we could imagine a hypothetical gene that could govern a calcium metabolism in bones, and produce a beneficial effect, such a rapid heal of them, would provoke, on the other hand, a steady deposition of calcium in arterial walls, and kill some older people. From the generalization of this idea, George C. William in 1957 coined the term antagonistic pleiotropy, that refers to the expression of a gene resulting in multiple competing effects, some beneficial but other detrimental to the organism. The pleiotropy stands from the Greek *pleion-*, means "more", and *-tropos*, meaning "way".

So the framework proposed by George C. William and Randolph M. Nesse introduces a good framework for our better understanding of the diseases. This ideas are supported in several studies, and we could perfectly trace those trade-offs at genetic level. For example, the mutated form of gene *ZIP4* that causes acrodermatitis enteropathica is surprisingly under selection in South-Saharian and Yoruban individuals (Engelken, J., et al (2014)). This autosomal recessive disease is caused by a deficiency in the absorption of Zinc ingested on the diet. It usually is characterized by marks in dermatitis, alopecia (loss of hair), among other phenotypes. But just to highlight the importance of zinc in our organism, as similarly to iron, has several functions in the body such as DNA repair, signaling, immune system and aging. Moreover, zinc dysregulation could produce diseases such as diabetes and cancer (Jansen J, et al (2009), Alam S (2012)). Why this ZIP4 mutation variant, concretely a Leu372Val substitution, could be under selection by an organism? The authors suggest that the prevalence in West African population was due to the reduction in intracellular zinc levels that may act by starving certain pathogens of Zinc. For this reason, despite this change to Leu to Val seems not to be favorable for the organism, have a good adaptation fitness in an environment rich in certain pathogens that are common in West African (South-Saharian and Yoruban people exhibit most extreme allele frequencies 0.99 and 0.96 respectively).

Under this prism, we see that the body has to deal with a packet of compromises, even after the process of maturation. In addition, genetic material has its own interest to propagate and replicate even at the expenses of the longevity and health

of a proper individual. And also that those little changes in genome could promote dramatic changes on the organism, sometimes for good, others for bad.

"A journey of a thousand miles begins with a single step"

Laosi

Index

1. INTRODUCTION

> "Let us understand what our own selfish genes are up to because we may then at least have the chance to upset their designs". Richard Dawkins

This thesis is focused into the study of some complex disorders through a relatively unexplored genomic variants of human genome. Concretely submicroscopic ancestral inversions. Two main models of complex diseases are studied: neurocognitive and immunological complex disorders. The aim of this introduction chapter is to present concretely those disorders and some of the relevant features of them, as well as why it is so difficult to untangle their complexity, compared for example with monogenic diseases. It also present a controversial concept missing heritability. That is most of genomic variants identified so far explain only a small proportion of the estimated heritability. Therefore where is it the missing heritability? The main aim of this thesis is try to found it in other genomic sources relatively unexplored such as events in mosaic and ancestral submicroscopic inversions.

1.1 Multifactorial studied disorders. Two models.

Multifactorial disorders are caused by a combination of genetic and environmental factors working together in ways that are not yet fully understood. Those disorders are also referred as complex disorders and it is thought than in most of the cases multiple genes are involved. Indeed it is more likely to be exceptional to found affected cases with few specific genes. Diseases may earn the label of multifactorial if they are clearly heritable and also influenced by environmental factors, or if the inheritance of some genetic risk factors is not sufficient to predict whether a person will actually develop the disease (Manolio et al (2009)). Besides the existence of environmental or stochastic factors, it is assumed that those disorders are polygenic. Therefore the terms 'multifactorial', 'complex' or 'polygenic' are commonly used as synonyms.

Examples of multifactorial diseases are asthma, diabetes, autoimmune diseases such as multiple sclerosis, cancers, heart disease, intellectual disability, inflammatory bowel disease, etc.

In opposition to multifactorial disorders one might distinguish from “Mendelian” or “single” (monogenic) gene diseases. In which, independently of the environment, it is enough to have a single or multiple key genes to produce that disorder. On monogenic disease (or Mendelian disorders) is more easy to detect those causative variants performing a linkage study. An illustrative example of monogenic disease discovered using linkage studies techniques is the Huntington’s disease (HD). HD is a neurodegenerative genetic disorder that affects muscle coordination and leads to mental decline and behavioral symptoms. It is caused by a mutated form of the huntingtin gene (HTT) of chr4, in which there is a CAG insertion with more than 36 repeats, when usual variants of healthy patients have a reduced number (6-35) (MacDonald, M. E. et al (1993)). HTT gene have an essential role as well in post-mitotic neurons and other cell types (with or without neural origin). In HD the trinucleotide repeat in HTT gene is translated as polyglutamine tail repeat in protein product provoking precipitations and proteic cumulus, specially in the brain. It is thought that this is the cause of this degenerative brain disease.

Usually when there is a severe deleterious effect over one critical gene, the usual outcome is the inviability of the organism. But sometimes, the organism survives at the cost to have some severe impairment. This is the case by rare disorders, and for this reason is possible to track the causative gene. I add in the supplementary material, as an author of this article, a case of monogenic disease in which three affected children present short stature due to growth hormone deficiency due to a pituitary hypoplasia (Jesús Argente et al (2013)). This case was caused by biallelic mutations in the gene RNPC3 gene. It encodes for a small nuclear ribonucleoprotein (snRNP) of the minor spliceosome system. A relevant point to have into account, is that the minor spliceosome acts over ~ 700 genes, and therefore, some of them were defective translated in affected patients. It was the first known case that a mutation of a protein of the minor spliceosome has been implicated in patients with isolated GH deficiency.

Multifactorial disorders do not usually have a single or several genes severely affected (at least not with the majority of the cases), otherwise the detection should be far easier. In multifactorial disorders is thought that the etiology usually is due to subtle changes in gene dosage, producing the disease. We could view the organism as an entity who tries to maintain the homeostasis, and this, is constantly threatened. When the organism fails

to respond effectively can result in disease or death. Disease could be viewed as a disturbance of homeostasis or steady state within an organism, while this homeostasis depends on the correct working of the critical genes as a whole system.

In summary, in multifactorial disorders, despite a significant genetic influence is found, no traits are 100% heritable. Furthermore, those traits show substantial environmental influence. The interaction (and correlation between genes and environment plays a significative role in development of complex traits (Plomin, R., & Deary, I. J. (2014)). The grade of genetic as well as environment influence could be demonstrated with twin studies. Examining the rates in identical (monozygotic) twins and fraternal (dizygotic) twins you could measure if a given trait have a strong genetic cause or not. If then a trait is 100% concordant (e.g. it is found in all twin pairs), you could consider that it is not significantly influenced by environmental factors. For this reason, it is difficult determine a person's risk of inheriting or passing these disorders.

Moreover, in multifactorial disorders, often presents a variable degree of phenotype-expression (expressivity), as well as variation in penetrance (if a sample carry a driving variant that also express the trait). Other observations points that multifactorial diseases sometimes occurs more frequently in one gender than in the other, but it is not a sex-limited trait. It also the disease could occur more frequently in a specific ethnic group than someone other (i.e., Africans, Caucasians, Asians, etc.). Besides this facts, diagnose correctly some of multifactorial disorders, as the neurocognitive ones, makes difficult to identify and classify correctly the affected samples based on their etiology observing the available phenotypes.

Some useful measures to analyze multifactorial disease are the following ones : grade of genetic and environment component, incidence, prevalence, and morbidity.

Incidence determines a person's probability of being diagnosed with a disease during a given period of time. Incidences shows how many new cases of a particular illness have been suffered by a community (or also how patterns of a condition within a population change over time). Therefore it informs about if it is decreasing or not given a concrete period of time.

Prevalence gives the person's likelihood of having a disease. It is the actual number of cases alive, with the disease either during a period of time or a particular date in time. Therefore number of cases in population.

Finally, comorbidity is the presence of one or more additional disorders (or diseases) co-occurring with a primary disease or disorder; or the effect of such additional disorders or diseases. Keep track of comorbidities for a given disease could help to better cluster patients to perform further studies.

During the last decades, the strategies tryied to identify responsible specific genes, do not succeed specially. Linkage analysis failed replicable linkage to chromosomal regions suggesting that they have little power to detect small effect sizes. But genome-wide studies (GWAS) studies also do not achieve the expected success. For example, no replicated genetic association account for more than one per cent of the population variance of quantitative traits such as height and weight (Plomin, R et al (2014)). We can conclude that there are no larger effect size, at least for SNP variants in those studies to date, as GWA studies have adequate power to detect those effect sizes. For example, in a study which includes nearly 18,000 children, no genome-wide significant associations were found, with a largest effect size that accounts for 0.2% of the variance of intelligence scores (Plomin, R., & Deary, I. J. (2014)). Other approximations were tried, such as Next Generation-Sequencing (NGS) methods that brights especially in detecting rare variants, but not on the common ones. Other approaches could involve the integration of several methods such as exome-sequencing and transcriptomic data, as it is in the case of Marta Codina-Solà et al (2015) (see supplementary information, as I participated in there as well).

We studied several neurocognitive and autoimmune-like multifactorial disorders. Concretely, the neurocognitive disorders was Autism Spectrum Disorder (ASD), Schizophrenia (SQZ), whereas autoimmune disorders, were Rheumatoid Arthritis (RA) and Colitis Ulcerosa (UC).

a) Neurocognitive disorders

a.1) Autism Spectrum disorders (ASD)

Autism spectrum disorders (ASD) [OMIM 209850] are a neurobehavioral group of syndromes characterized by deficits in social interaction, impaired communication skills, repetitive, stereotypical and ritualized patterns of behavior and interests. Is cataloged as neurodevelopmental disorder, which phenotypic traits typically appear before the age 3 years old.

There is no common agree of the prevalence of ASD, and in the last decades, the estimated proportion was 4/10,000 to 1/110 children, but seems to be an increasing incidence in the last years (Centers for Disease Control and Prevention (2012),Report, M. W. (2014)). Still, there is a debate on how much this increase is due to diagnostic improvements, or an emerging environmental factors (Marco, E. J. et al (2006)).

The clinical diagnostic of ASD is still not so clear, making it as a broad phenotype rather than as a single clinical entity. Large efforts have been made in order to create instruments for diagnosing autism in children, and particularly there are two that are widely used the Autism Diagnostic Observation Schedule-Generic (ADOS-G) and Autism Diagnostic Interview-Revised (ADI-R). Both tests are based mainly on the behavioral criteria listed in the 1994 by American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV). DSM-IV compiles the general features of affected samples based mainly on the three symptomatology behaviours: qualitative impairment in social interaction, qualitative impairments in communication and restricted repetitive and stereotyped patterns of behavior, interests, and activities. Based on each features, ADI-R and ADOS-G propose different tests to diagnose autism. ADI-R and ADOS-G seems to be the gold standard in the diagnostic process of ASD and the results shows that approximately 75 percent agreement with team diagnosis (De Bildt (2004), Mazefsky, C. A. & Oswald, D.P. (2006)). Besides that, there is a common agreement on which social and communication impairments are key diagnostic characteristic of autism (DSM-IV and DSM-V), although there is still a room for improvement.

Hypothesis and pathophysiological data

Some psychological studies, highlight the role of language (in the widest sense) in the creation of the so called theory of mind (Hale, C. M., & Tager-flusberg, H. (2003)). Theory of mind supports the fact that in healthy individuals, there is an ability to understand another person's mental state, beliefs, intents and desires, as separate from one's own thoughts, experiences and behaviors (McGovern, C. W., & Sigman, M. (2005)). In the individuals with autism, on the other hand, they do not present this ability, and it also could explain their delays in language acquisition. It is estimated that 30 and 50% of individuals with autism never develop functional speech, and among those that develop verbal communication, it is often restricted to expression of instrumental functions, or simple labeling (Hale, C. M., & Tager-flusberg, H. (2003), McGovern, C. W., & Sigman, M. (2005)). Research highlights the importance of joint attention (that is alerting another with nonverbal stimulus, typical starts to appear in infancy, during the first year of life) in predicting the language and communication deficits that are the hallmarks of autism (McGovern, C. W., & Sigman, M. (2005)). Over the past decade, some researchers have proposed that mirror neuron dysfunction might underlie the behavioral manifestations presented in autism (Iacoboni, M., & Dapretto, M. (2006), Oberman, L. M. et al. (2008)). Imitation is the widely used form of learning during development, and also the central to the development of fundamental social skills (e.g. as reading facial and other body gesture and for understanding the goals, intentions and desires of other people). In ASD seems that there is a dysfunctional or broken mirror neuron system (MNS) (Iacoboni, M. et al. (2006), Oberman, L.M. et al (2008)). Mirror neurons are involved not only in the perception and comprehension of motor actions in humans (Gentilucci, M. et al (2000), Iacoboni, M. et al (2006), Perra, O. et al. (2008)), but also a higher-order cognitive process such imitation and language (Rizzolatti, G. et al (1996), Iacoboni, M. et al (2006), Oberman, L.M. et al (2008)). In fact those neurons are found in Broca's area (Brodmann area 44), an area of the inferior frontal cortex that has been strongly linked with language, as well as other areas such as the inferior parietal lobule and superior temporal sulcus as shows different studies performed using fMRI (functional magnetic resonance) (Gentilucci, M. et al (2000), Buccino, G. et al (2004), Iacoboni, M. et al (2006)). Therefore, seems to be common agreement that language, motor-skills and social seems to be related.

Indeed, regarding motor-skills, there is a study of Perra, O. et al (2008) that shows impaired performance in both gestural imitation and general motor skills in autistic individuals. Regarding some studies promotes certain criticism against the theory of mirror neuron system hypothesis and their role with autism (Hamilton, A. F. D. C. et al (2007)), there is a major trend in supporting it in ASD (Yang, D.Y.-J. et al (2015)).

Among the already commented brain functional features, using voxel-based morphometry (VBM), it is also observed a decrease of gray matter density in Brodmann area 45 in the left inferior frontal gyrus in adults with autism (De Fossé, L., et al 2004). Other studies using MRI, shows that there is a brain size increase during the first 3 year of life in autistic child, and also that neurons are more packed (and in fact there are an increased number and more glial density in prefrontal cortex) and a reduced number of Purkinje cells in the cerebellum (Muotri, A. R. (2015) review). Also, other studies points altered cortical thickness in autism patients Chien, H,.Y. (2015).

Genetic basis

The ASD is more common in males than females, with at least a ratio of four times higher, and also the prevalence varies among racial/ethnic group (Miles, J. H. (2011), Report, M. W. (2014)). Evidence suggests that genetic influences from X chromosome or inherited epigenetic markers (inherited from mainly from fathers), could have an influencing in engendering this male vulnerability (Marco,E.J et al (2006), Crespi N., and Badcock, C. (2008)). Some of the X-linked genes come from aneuploides (Turner syndrome and Klinefelter syndrome), trinucleotide expansions (Fragile X syndrome) and nucleotide mutations (Rett Syndrome (X-linked dominant mutation in MECP2 with lethality in hemizygous male), Neurologins 3 & 4, and SLC6A8).

There is a high percentage of heritability in ASD (about 90%) given the concordance rates observed in monozygotic twins studies, and siblings have an approximately 50-fold increased risk of ASD. This makes autism one of the most heritable neuropsychiatric disorders.

The etiology of the majority of the cases (~70%) remains unknown. The genetic studies performed till now, not succeed in finding common variants. In fact, very few common variants are found in ASD that will be replicated in different datasets. The largest meta-analysis study in 2012 highlight those 6 SNPs (Klei, L. et al (2012)) :

SNP	OR	pvalue
rs4307059	1.19	$2.1*10^{-10}$
rs7704909	1.17	$9.9*10^{-10}$
rs12518194	1.16	$1.1*10^{-9}$
rs4327572	1.15	$2.7*10^{-9}$
rs1896731	0.87	$4.8*10^{-8}$
rs10038113	0.87	$7.4*10^{-8}$

Table 1.1. Most significative SNPs in common in ASD (Klei, L. et al (2012)).

This fact suggests that could be for the high degree of genetic heterogeneity in ASD, and maybe the sample size performed in most of the studies is relatively small. Indeed, recent studies highlight the importance to characterize better the samples in ASD (Chaste, P. et al (2015)). For example, in our results, as well as other observations (Klei, L. et al (2012)), we noticed that the architecture of simplex and multiplex families could be different in autism. A recent study on simplex families focus in subsets of ASD found sensible high OR (~2.0) (Chaste, P. et al (2015)).

Regarding point mutations and CNV studies, molecular karyotypic shown that 5 % to 10 % of patients carry chromosomal rearrangements, and the burden of rare, large (>1Mb) and de novo copy number variants (CNV) is higher among ASD patients than controls (Sanders SJ et al (2012), O'Roak BK et al (2012), Neale VM. et al (2012), Codina-Solà, M. et al (2015)) . That studies show shows hundreds of genes implicated in the genetic etiology of ASD, with few of them recurrently mutated in unrelated cases. Most recurrently mutated genes are related in synaptic connectivity. Concretely, the neuroligin (NL) gene family family that codes for brain specific cell adhesion molecules appear to be linked with ASD cases (Gutierrez, R. C. et al (2009), Miles, J. H. (2011)). Those genes, such as neurexin 1, neuroligin 3 & 4, SHANK 3 (glutamatergic synapse abnormalities in ASD), R471C-NL3, NLGN4Y, CTNND2 (adhesive junction-associated

d-catenin protein) will support the hypothesis of altered synchronicity both within and between brain regions (Gutierrez, R. C. et al (2009), Pinto, D. (2010), O'Roak, B. J. et al (2012), Turner, T. N. et al. (2015)). Other genes that appear recurrently mutated CACNA1C, CACNA1F also seems to be related with a calcium channel disorder characterized with severe QT prolongation, syndactyly, cardiac defects, dysmorphic facies, mental retardation, developmental delays, and autistic symptoms (Miles, J. H. (2011)).

Other set of genes, points on a dysregulation of serotonin and/or its regulatory networks (hyperserotonemia is another old-known biomarker in ASD), particularly mGluR5, seems also to have an important role in autism (Veenstra-VanderWeele, J., & Blakely, R. D. (2012)).

Other findings suggests that the etiology could be related for the packaged neurons observed. For example, CNV 16p11.2, in which CUL3-GTPase RhoA, seems to be a major regulator of the actin cytoskeleton and cell migration (the RNAi knockdown of CUL3 leads to accumulation of actin stress fiber formation and impaired cell migration (Chen et al (2009), Lin, G. N. (2015))).

Finally, other studies also hypothesize that the triggering in ASD in some cases could be related with the immune system (Ashwood, P. et al (2006), Rossignol, D. a, & Frye, R. E. (2012), Mead, J., & Ashwood, P.(2015), Pramparo, T., et al (2015), among many others). The findings that supports those hypothesis, is that there is a significative signature of differentially coexpressed genes related the immune and inflammatory functions genes (and also mitochondrial dysfunction and oxidative stress) found in most of the case-control performed studies (Rossignol, D. a, & Frye, R. E. (2012)). In fact these findings also motivates the idea to find biomarkers of ASD in immune system (Ashwood, P. et al (2006), Pramparo, T., et al (2015)). Moreover, it has been reported in autistic children abnormal or skewed T helper cell type 1 (TH1/TH2) cytokine profiles, decreased lymphocyte numbers, imbalance of serum immunoglobulin levels. This could be consequence or cause of the disorder? Further investigations are required.

In any case, the global pathway of autism is not yet understood, and aetiological heterogeneity of ASD certainly does not certainly help. And indeed, one may have to

account that the architecture of simple and multiplex autism would be strikingly different (Sebat, J., et al (2007) and our own studies on ASD, see results sections). But still, as the majority of the studies performed shows that the majority of liability percentage will reside in common variation, but this portion is uncovered. The main hypothesis is that individually of small effect, have a substantial impact en *masse* (Klei, L., Sanders et al (2012), Gaugler, T. et al (2014)).

a.2) Schizophrenia Spectrum disorders (SSD)

Schizophrenia spectrum disorders (SSD) [OMIM 181500] are a set of psychiatric disorders characterized by some of those clinical symptoms : auditory hallucinations, delusions, and disorganized speech, social withdrawal or cognitive dysfunctions. The etymology of schizophrenia comes from Greek *skhizein*, meaning "to split", and *phren*, meaning "mind", altogether the term means "splitting of mental functions" (instead of "split personality" or "multiple personality disorder" as a common misconception in public perception). The core of this disorder is the poor functional integration of sensory, cognitive and affective process, also called MSI or Multisensory Integration.
Unaffected MSI individuals are able to create a spontaneous perceptual-cognitive process that with relevant information from multiple sensory modalities are capable to generate a holistic experience. On the other hand, individuals with deficit in MSI, lacks an appropriate behavioral response in a complex environment (Tseng H.H. et al (2015)). Indeed, available evidences indicates an impaired MSI for non-emotional stimuli especially for linguistic information, in schizophrenia.

Schizophrenia has biological basis as several studies points out that there is heritable. Concretely is found that there is a concordance of 40-60 % in identical twins, whereas in fraternal twins is about 10-20 % (Gottesman and Shields(1982), Gottesman (1991)). This suggest a complex etiology influenced by genetic and nongenetic factors. Their symptoms begin typically in young adulthood, and about 0.3–0.7% of people are affected during their lifetime.

Pathophysiology

Although exists pharmacological treatment for schizophrenia, namely antipsychotic drugs , their efficacy is poor for many patients. Nevertheless, when a general treatment works against some disease, although it do it partially, it could help to understand better the underlying architecture of the disease. This is the case of the core treatment for SSD are based on a mechanism discovered over 60 years ago. It is based on the blockade of the type 2 dopaminergic receptor (Seeman, P. et al (1975)). Despite to be effective against psychotic symptoms, have high rates of neurologic side effects, such as extrapyramidal signs and tardive dyskinesia. Since there, no new antipsychotic drug of proven efficacy have been developed (try other agents with lower affinity for dopamine D2 receptors and greater affinities for other neuroreceptors, such as serotonin and norepinephrine) (Ripke, S. et al. (2014)). Other more recent discovered treatment was the case raloxifene, but just as an adjunctive treatment (Weickert, T. W. et al (2015)). Raloxifene seems to improve attention and memory in men and women with schizophrenia . This treatment is based on the role that estrogen hormone could have in Schizophrenia, based on previous studies that notice the asymmetry between the women and men first-episodes of psychosis. That is observed that women have a greater likelihood of developing first-episode psychosis when estrogen levels are low (during an estrogen trough in the menstrual cycle, post partum and around menopause). Whereas schizophrenia symptoms can remit during pregnancy when estrogen and progesterone levels are high (Häfner, H. (2003)). Those observation motivates the creation of raloxifene treatment. Raloxifene stimulates estrogen-like activity in brain and can improve cognition in older adults, and also shows some samples with carryover effects on cognitive benefits even after raloxifene withdrawal (Weickert, T. W. et al (2015)).

Regarding physiological features related in schizophrenia, some studies notice gray matter reductions in prefrontal and temporal cortices (Cooke et al. (2001), Bergé et al (2011)) in patients with insight deficits. Additionally, they present cortical thinning in the middle, inferior and medial frontal gyri, temporal gyri and precuneus as well as significant correlation between insight deficits and gray matter reduction in cerebellum.

On the other hand, functional studies on brain of affected samples with active psychosis, ratifies and add more evidence on the already observed fact that a lower ratio of anterior to posterior cerebral activity in patients with schizophrenia (Fu, C. H. Y. et al (2014)).

This concept was already observed on 1974 by Ingvar DH and Franzen, and purpose and coin the “hypofrontality” hypothesis among the schizophrenic samples.

Genetics basis

GWAS analysis, shows, as well in ASD, hundreds of loci associated with SSD. Recently, it was created a Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC), and the last report effectuated combining all available schizophrenia samples with published or unpublished GWAS genotypes, recorded a highly suggerent 108 loci (Ripke, S. et al. (2014)). Summarizing this work, they confirm the major hypotheses of the aetiology and treatment of schizophrenia DRD2 and many genes (GRM3, GRIN2A, SRR, GRIA1) involved in glutamatergic neurotransmission and synaptic plasticity. Association of several genes that encode voltage-gated calcium channel subunits, such as CACNA1C, CACNB2 and CACNA1I, are also confirmed. In addition, rare variation genes encoding calcium channels, and proteins involved in glutamatergic neurotransmission and synaptic plasticity. Several of them, as commented in previous section, are also in common with ASD.

Other subset of genes are somehow related the immune system (Benros, M. E. et al. (2012), Ripke, S. et al (2014)), that could be applied in subgroups of patients with schizophrenia. Those, such as genes seems to play an active role on important immune functions, and supports the hypothesis of inflammatory mechanisms that will affect the brain will trigger the schizophrenia disease. Those results confirm the clinician's observation in 1950s and 1960s,that seemed to be an unusually high occurrence of celiac disease in persons with schizophrenia, in a subpopulation with some sort of vulnerability (about 5 times many celiac persons affected with schizophrenia than in general population).

Other features observed in postmortem brain studies of schizophrenics (and in animal models of schizophrenia), is that shows a decreased expression of the GABA synthetic enzyme glutamate decarboxylase 67 (GAD67) in a subset of GABAergic neurons (Fujihara, K. (2015)). Other studies performed with Electroencephalography (EEG, that is a non-invasive tool to record electrical activity of the brain), shows an abnormal function of GABA neurons could produce an impairment on network oscillations and trigger Schizophrenia (Gonzalez-Burgos, G. (2008), Fujihara, K. (2015)). This could be the phenotypic effect of the decreased expression of GAD67.

Environmental risk factors that have recognized to play a role in the etiology of schizophrenia includes place and season of birth, maternal obstetrical complications, parental age, neonatal vitamin D levels, prenatal infection (eg, influenza, toxoplasmosis, and herpes simplex virus, and important weight (OR 8.1; 95% 3.24-20.3) on low social class (Agerbo, E. et al (2015)). But again, regarding epidemiological studies and with the exploration of environmental risk factors, is that schizophrenia is a heterogeneous disorder that varies across race/ethnicity, sex and age and place. Indeed, it does not exist a gold standard to diagnose schizophrenia, and the criteria may vary between different places. In previous study (Agerbo, E et al (2015)), grouping all markers that individually does not achieve significance in a large-scale association study, also called polygenic risk score (PRS), achieves an OR 8.01; 95% CI, 4.53-14.16 on Danish population (871 control individuals vs 866 cases). In that paper they also calculated a polygenic risk (PRS), that is the sum of schizophrenia known risk alleles carried by an individual, weighted by the respective ORs, as a predictive score. Nevertheless, PRS values showed similar effect (slightly less) than social class. It also have been into account, that the PRS captures only a proportion of the variation attributable to common SNPs, and it is not expected to capture deleterious exonic mutations or rare genetic and copy number variation. The value of PRS suggest how little we know about the biological pathway between those alleles and schizophrenia, and reveals that an important fraction of missing heritability still exist.

a.3) ASD and SSD separate entities or pieces of same puzzle?

Remarkably, ASD and Schizophrenia share multiple phenotypic similarities and risk factors, and have been reported to co-occur at elevated rates (Chisholm, K. et al (2015)). The overlap between ASD and SSD includes phenomenological, genetic, environmental, and imaging evidence. Historically, was believed that autism was a central feature of schizophrenia, and in fact the term autism was used interchangeably with schizophrenia until the 1970 (Chisholm, K. et al (2015)). As SSD has a typical adolescent onset with prominent positive psychotic symptoms, while ASD is characterized by deficits in social interaction, communication and behaviour that begin within the first few years of life, the nosologic[1] separation appear to be justified. Nevertheless, SSD and ASD share multiple phenotypic similarities and risk factors (Hamlyn, J. et al (2013), Annelies A. Spek et al (2010)). Both disorders include deficits in social interaction and communication as primary symptoms. Lack of emotional in ASD can be compared to a lack emotional response in SSD. Delay or lack of speech development in ASD in poverty of speech matches as well is SSD, and catatonic features are observed in both diagnoses. Moreover, both have been reported to co-occur at higher rate than expected in population, and both are neurodevelopmental rather than neurodegenerative disorders. In populations with ASD, rates of comorbid SSD have been variably reported from 0% to 34.8% (Chrisholm, K. et al (2015)).

At genetic point of view, the heritability on both disorders are quite similar around 50-80%. And, remarkably, although autism and schizophrenia are considered two different neuropsychiatric disorders, recent studies indicate that they share genetic factors. An example of illustrative gene, was contactin-associated protein-2 (CNTNAP2) that encodes a protein of neurexin family and have some role in vertebrate nervous system as cell adhesion molecules and receptors, and it is associated with potassium channels (Burbach, J. P. H. et al (2009)). CNTNAP2 is regulated by FOXP2, and CNTNAP2 polymorphisms are detected in children with specific language impairment (Vernes, S.C. et al (2015)).

[1] Nosology (from Ancient Greek νόσος (nosos), meaning "disease", and -λογία (-logia), meaning "study of-") is a branch of medicine that deals with classification of diseases. Diseases may be classified by etiology (cause), pathogenesis (mechanism by which the disease is caused), or by symptom(s).

Other structural variants, such as copy number variants (CNVs), shows that specific rare alleles have been found to occur in both disorders (Chrisholm, K. et al (2015)).

Moreover, a study performed by Carroll, L.S., & Owen, M.J. (2009) in which we try to view the genetic differences at genetic level among autism, schizophrenia and bipolar disorder, they were astonished with the similarity of results between ASD and SSD due the high number of shared CNV duplications and deletions, such as 16p11.2, 3q29, 7q36.3, 15q11,17p12, 22q11.2, 1q21.1 and 15q13.3 (Kim, Y. S., & Leventhal, B. L. (2015)). Those CNV are related with intellectual disabilities, developmental delay, speech problems, schizophrenia, seizures, increased body weight or obesity, and increased head circumference .

Other studies that focus with brain activity, it was found abnormal neural oscillations and synchrony in schizophrenia and autism (Gutierrez, R.C. et al (2009), Rosin, B., et al (2011), Uhlhaas, P. J.,(2010), Gonzalez-Burgos, G., & Lewis, D. a. (2008)). Concretely, autism, alzheimer's, and schizophrenia are associated to lack of synchronization of neurons, while others brains disease such as epilepsy and Parkinson's are associated with excessive synchronization.

Regarding environmental factors SSD and ASD share a remarkable number of environmental factors, most of them related to obstetric (science that is focused on pregnancy, childbirth and postpartum period , literally in latin *obstare* means the waiting time) complications (Hamlyn et al (2013)). Also the paternal age has also been identified as potential risk for both ASD and SSD (Gardener et al, (2009)) , and particularly paternal age of > 50 (1.24 [30-39 vs <30]; 1.44 [40+ vs 25-29]). Other studies suggest that maybe, exploring epidemiological data, that Schizophrenia and Autism, could share a common pathogenesis via perinatal inflammation (Meyer, U. R. S. et al (2011), Ornoy, A. et al (2015)).

All those findings raise the possibility that those neurocognitive disorders share pathogenic mechanisms and that similar defects in biological pathways of brain development might explain the phenotypic spectrum of this disorders. More details of ASD and SSD were briefly described in following sections.

b) Autoimmune multifactorial diseases

Autoimmune diseases are illnesses that occur when the body's tissues are mistakenly attacked by their own immune system. It is thought that the causation are based on the formation of autoantibodies. Clinicians differentiate the autoimmune disease according if they are organ-specific (e.g. diabetes mellitus) or if the disorder involves autoantibodies that are not specific to antigens found on certain tissues or systemic (e.g. systemic lupus erythematosus, rheumatoid arthritis, etc.).

The pathology mechanism is still unknown, but just to have a brief global idea of how autoantibodies are built is the following. Antibodies are produced by B cells in two ways: randomly and in response to a foreign protein or substance within the body. In those affected samples, an autoantibody is produced by B cell that is directed against one or more of individual's own proteins.

The initiation of this autoreactivity could come mainly from environmental factors, infectious agents, and noninfectious (Ian R. Mackay M & Fread S. Rosen (2001)). That is, besides the genetical predisposition is also necessary an environmental exposure or a change in the internal environment for autoreactivity. Exposition to pathogens (and their microbial or viral antigens) have the potentiality to initiate the autoreactivity process. Many autoimmune disease are more frequent in women than in men. Indeed estrogens increase the severity of systemic lupus erythematosus. And other noninfectious elements that could trigger the disease could be for example drugs.

When we talk about multifactorial immune disease in inevitable to talk about major histocompatibility complex (MHC). MHC contains a set of molecules involved in immune recognition and signalling between immunity cells. The name *histocompatibility* cames from *histo-*, that means "tissue" and the part of compatibility represents a duality between the self and nonself. MHC is a region situated on the short arm of chromosome 6 that harbors the largest number of replicated associations across human genome for a wide range of diseases, although functional basis for these associations are poorly understood. MHC comprises a contiguous ~4 Mb region on short arm of chromosome 6 (6p21.31), the extended MHC (xMHC), spans an even larger 7.6 Mb region and comprises more than 400 annotated genes and pseudogenes.

Due that is the classical human leukocyte antigen loci, is also called HLA 6 (De Bakker, P. I. et al (2012)). Is characterized to be a very hypervariable region, among samples (and within samples if leukocytes cells are genotyped!). HLA alleles are strongly differentiated across global populations, and therefore highly sensitive to type 1 error when slight differences in ancestry of cases and controls are not adequately controlled for.

Just to have a big picture of the systemic way in which some MHC genes work, and refresh some basic key concepts of adaptive immune system without pretend be very exhaustive, but to see how it works on a systemic perspective. I will focus only on the main interaction of T-cells, B-cells, macrophages and pathogens as the main actors. There are different kinds of T-cells: T helper cells (interact with memory B cells and also activate cytotoxic T cells and macrophages), cytotoxic T cells (that destroy virus-infected cells and tumor cells), memory T cells (that have a subset of antigen-specific T cells that persist long-term after infection or vaccination) among many others. If certain pathogen is detected by T cell, an immune system response is triggered, and the pathogen eliminated. In some cases, T cell does not detect that pathogen. In those cases, it is likely that could be detected by the some B-cells with its specific antibody. B-cells are continuously produced in bone marrow, and they have its specific antibody. When one of the B-cells matches with a given antigen, B-cell engulfs the antigen and digest it. Then, it displays the antigen fragment bound to unique MHC molecules. This combination of antigen and MHC attracts to help of mature matching T-cell. Then cytokines are secreted by T-cell produces B-cell to multiply and mature into antibody. Then, those antibodies were released into the blood, and lock onto matching antigen. A a result, those complexes with antigen-antibody, were cleared by complement cascade.

In following figure there is a basic scheme of how it works this subsystem.

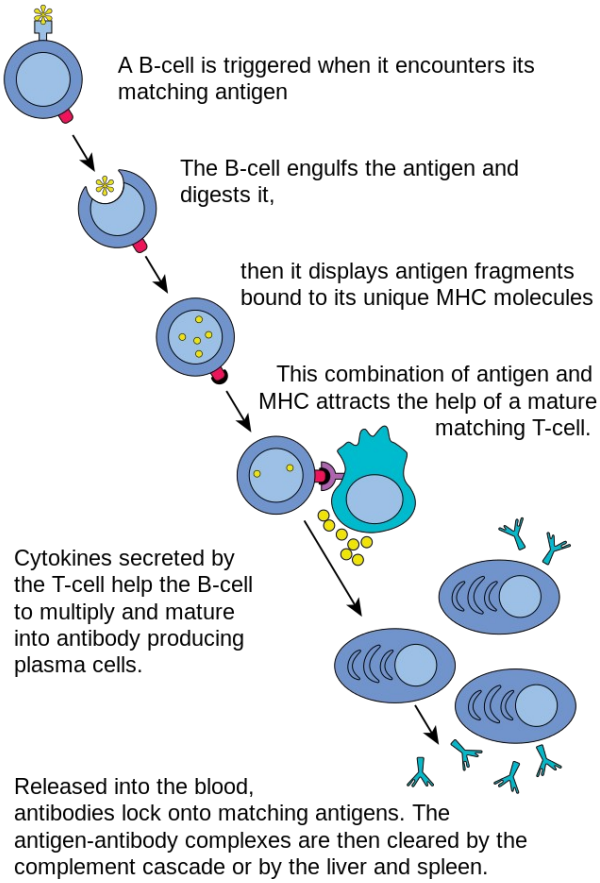


Fig 1.4 Visual schema of "B cell activation" by Fred the Oyster (Licensed under Public Domain via Wikimedia Commons)

In this process, we can have a closer look to some of the point that is one of the major responsible of the injury on most of autoimmune disease. In a healthy individuals, cytokine signals cause B-cells to multiply, and some of these B-cells turn into plasma cells that secrete antibodies (immunoglobulins). These antibodies then circulate into the bloodstream so that when they encounter the antigen again, they bind into it, forming a complex that is then acted on by other cells of the immune system into an effort to destroy the invader. The side-effect is that this process causes some inflammation and injury on healthy tissue, but usually, the immune system itself, controls this inflammation.

Certainly it is not a coincidence that most of the immune disease, have its principal suspects among MHC genes. For example, in lupus erythematosus and other autoimmune disease are based on causes that the immune system targets own body cells

for destruction. Affected samples usually presents an overactivation of B cells and T cells that promote an over-expression of cytokines on tissue affected, generating more inflammation and tissue injury (Dean, G. S. et al (2000)). This increased inflammation is the cause of the pain and discomfort, but the major problem is with potential long-term irreversible scarring underlying the disease.

Lot of effort is centered on revealing the underlying mechanisms of the most common immune disease. Particularly we are trying to contributing our grain of sand to understand little more Rheumatoid Arthritis (RA) and Ulcerative Colitis (UC) multifactorial diseases.

b.1) Rheumatoid arthritis RA

Rheumatoid arthritis (RA) is an autoimmune disease characterized by persistent synovitis, systemic inflammation and autoantibodies. The term arthritis came from Greek *artr-*, that means articulation, and *-itis*, that means inflammation, therefore arthritis cames from the inflammation of a joint (articulation). On the other hand, rheumatoid cames from *rheuma/rheumatos* (Greek) that means that something flows(or it is fluid, current), and the suffix *-oid* means reassembling, altogether Rheumatoid arthritis could means something as *joint inflammation* that resembles *rheumatic fever.*

Genetic factors account for 50 % of the risk to develop RA, and the main environmental risk lies on smoking. In industrialised countries, RA affects 0.5-1.0% of adults, with 5-50 per 100,000 new cases annually. The genetic component has long been established, and twin analysis shows the heritability to be about 60% (Amos C. et al (2006)). RA is most typical in women and elderly people, and researches believe that there may be a hormonal component as many of the autoimmune conditions are much more prevalent in women in childbearing age (Scott, D. L., Wolfe, F., & Huizinga, T. W. J. (2010)). Uncontrolled RA causes causes joint damage, decreased quality of life, disability, and cardiovascular and other comorbidities. A schema could be shown in Fig 1.5.

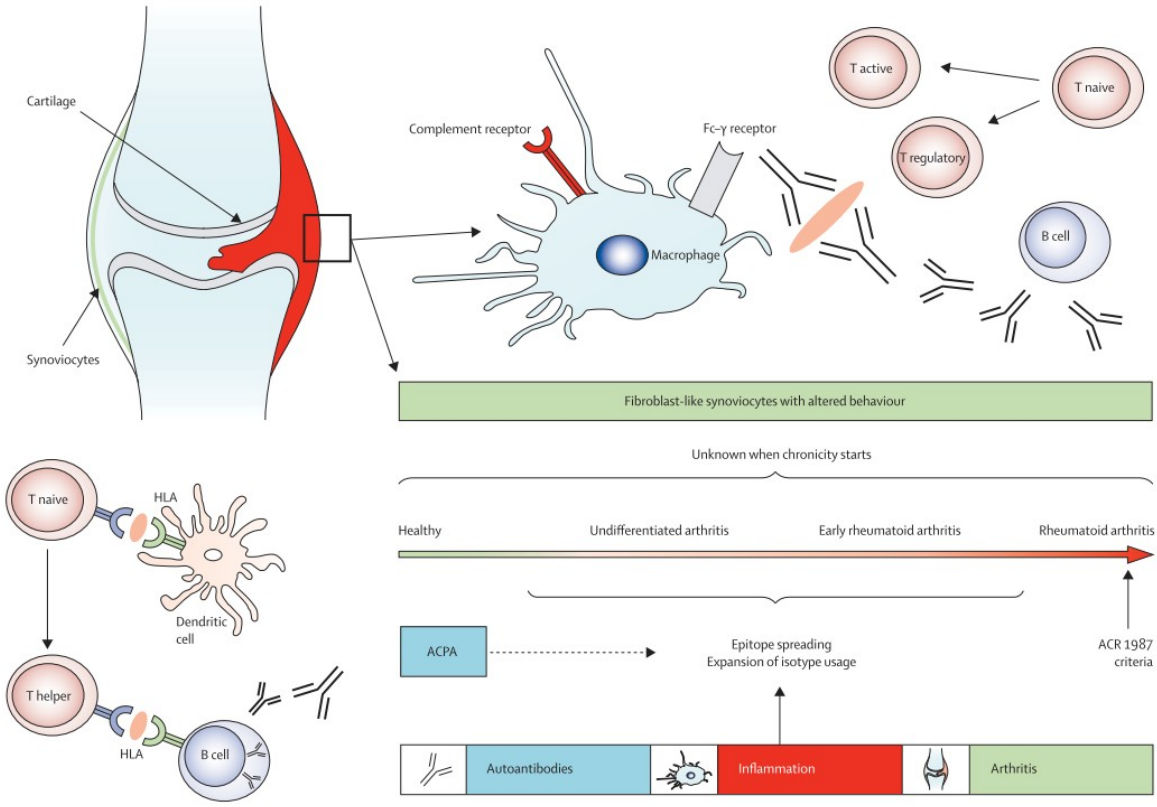


Fig 1.5 Schema of RA affected joint, in which synovial membrane is inflamed due a fibroblast-like and macrophage-like synoviocytes, macrophages and several populations of B and T cells. Macrophages produce all kind of proinflammatory products (e.g. TNF) such as immune complex of union to Fc-γ receptors or complement receptors on their surface. In this schema also shows ACAPs (Anti–citrullinated protein antibodies) as the main responsibles to lead to RA disorder (Scott, D. L., Wolfe, F., & Huizinga, T. W. J. (2010))

Genetics and treatment

RA is considered a not just a single disease so a clinical syndrome spanning several diseases subsets. This different subsets contain their own inflammatory cascades, which all lead toward the final common pathway that provokes damage to articular cartilage and the underlying bone. One of those key elements in one of the more common cascade, includes overproduction and overexpression of TNF. This overexpression of TNF could have several causes, such interactions between T and B lymphocytes, synovial-like fibroblasts and macrophages. This process leads to overproduction of many cytokines (such as interleukin 6), which drives to tissue inflammation and joint destruction. Therefore the regulator of this cascade became a target to treat this immune disorder.

RA are treated basically with the so-called Disease-modifying antirheumatic drugs (DMARDs), that reduces synovitis and systemic inflammation. The leading DMARD is methotrexate, that is combined with other drugs. In some cases also biological agents are used when arthritis is uncontrolled, as for example tumor necrosis factor inhibitors. Usually the finding of key biomarker profiles are the key element to long-term remission.

b.2) Ulcerative Colitis UC

Ulcerative Colitis is subcategory of inflammatory bowel disease that causes inflammation and ulcers in the colon. Colitis comes from *-itis*, that in Greek means inflammation (or infection), and Ulcerative comes from the characteristic ulcers that appear in the colon. Colon is the largest section of large intestine, and the main symptom is a constant diarrhea mixed with blood that usually is confused with Crohn's disease or irritable bowel syndrome. Although UC shares much in common with Crohn's disease, as UC names suggest, is af infection only affects the colon and rectum, while Crohn may affect any part of the gastrointestinal tract, from mouth to anus.

UC have an incidence of 1.2 - 20.3 cases for every 100,000 persons per year, with a prevalence of 7.6-246 cases per 100,000 per year (Colitis, U. (2003), Danese, S., & Fiocchi, C. (2015)), with varies considerably among population. UC is an intermittent disease, with periods of exacerbated symptoms, and periods that are relatively symptom-free.

The exact cause of UC remains undetermined, and the condition, as well as RA, is somehow related to a combination of genetic and environmental factors.

Genetics and treatment

As commented, there is a clear asymmetry in prevalence, e.g. in european population is rare (less than 1%), whereas in Japanese population it is highly prevalent (20%-25%). The most reproducible association observed is HLA-DRB1, but this variant precisely it

has low prevalence to europeans (Fernando, M. M. a et al (2008)). In any case, most of the heritability of UC (>70%) has still not been characterized.

The usual genes implicated in UC are related with the dysfunction of epithelial barrier (ECM1, HNF4A, CDH1, and LAMB1); apoptosis and autophagy (DAP); transcriptional regulation (PRDM1, IRF5, and NKX2-3). In addition, multiple genes in the interleukin-23 signaling pathway overlap in UC and Crohn's disease (e.g., IL23R, JAK2, STAT3, IL12B), as well as genes of the HLA-DR (Danese, S., & Fiocchi, C. (2015)).

Additionally, it has been observed that composition of the gut microbiota, and/or defects in mucosal immunity, could lead to UC, although that more evidence is needed (Colitis, U. (2003).

Conventional treatments usual are effective in maintaining remission and decrease the length of active disease periods, but unfortunately with side effects, and a significant people fail to respond to most of the drugs. Usually are based on the interleukin-13-blocking agents (interferon beta-1a). Depending on the severity of the disease, and could involve the surgical removal of the abnormal part of the colon and rectum, and connect the healthy ends.

Environmental factors, such smoking, diets high in fat and sugar, stress, medication use and high socioeconomic status are the most influential on the appearance of that disorder.

1.2 Genetics of multifactorial disorders: the missing heritability

One late evening Nasreddin found himself walking home ...
"Mullah, please tell me: What is wrong?"
"Ah, my friend, I seem to have lost my keys. Would you help me search them? I know I had them when I left the tea house."
So, he helps Nasreddin with the search for the keys. For quite a while the man is searching here and there but no keys are to be found. He looks over to Nasreddin and finds him searching only a small area around a street lamp.
"Mullah, why are you only searching there?"
"Why would I search where there is no light?
Mullah Nasreddin Tale

The studies performed on identically genetically individuals (monozygotic twins) and pedigree family of affected cases make possible to calculate the heritability proportion on some multifactorial disorders, and measure the genetic component. For example, in the case of ASD the heritability is due around of 65-80% -see section of Neurocognitive disorders for more detail. Despite this high proportion of genetic component, most of the GWAS performed studies points that there is still high proportion of heritability that it is still uncovered, and they refer them as the missing heritability (Manolio, T. et al (2009),O. Zuk et al (2012)). Most of SNP variants found in GWAS studies, explains small increments in risk, and explains a small proportion of the full heritability. Some studies supports the idea of that the explanation could be that some of common variants are infravalorated (as suggested by Walter Bodmer and Carolina Bonilla (2008), Manolio, T. et al (2009), Elizabeth T. Circulli et al (2010)). Several studies are also followed this approach (f.ex Hrein Stefansson et al (2009), Jianxin Shi et al (2009), among many others).

Therefore, with large sample sizes of affected and unaffected samples is possible to trace trails of genetic component. The rationality under this strategy in the multifactorial disease is that if the component explained by genetics is high, despite there could be several genes implicated, if we have enough sample size, we will be able to found the positive signals. And, the other way around also work, even better, if the genetic component is very low, but the genomic variant is a single gene, it could be (usually)

easily found. Generally, as high are the genomic variants implicated, more sample size is required according to this hypothesis. Nevertheless, despite the enormous efforts performed with GWAS studies, the genetic basis of common, multifactorial disorders remains poorly understood. Following two decades of research with the traditional candidate-gene approach, very few genetic regions and genes can reliably be considered true positives. The majority of gene-disease association findings have shown inconsistency and non-reproducibility (Kistios, G. (2010)).

In the previous section we assessed that the mutation is a crucial component of evolution, and inherent to the evolution system. The effect on that mutation relies on its size and the loci affected. Despite this fact, the mutation *per se* is not an event usually welcome by the body, and it has its own mechanisms to avoid such chromosomal aberrations during the formation of a new organism, so that is producing spontaneous abortion (Qumsiyeh, M. et al (1999)), but sometimes fail producing those diseases.

According to the size, we can distinguish those small-scale mutations, such as the point mutations, that provokes an exchange of a single nucleotide for another. Or those large-scale that comprises CNVs, mosaic events and inversions. All that variations were commented in following sections.

a) Rare variation

Rare variation is all those chromosome mutations that are uncommon events in our genome that could be main responsible of a given disorder. We present those ones based on their affecting area size.

Point mutations

Point mutations, could, at the same way, be classified as silent mutations, missense mutations and nonsense mutations. The silent mutations are those changes in the nucleotide that, when the final codon was translated, will codify for the same (or very similar) equivalent amino acid. The missense mutations, in which codify to a different amino-acid. Nonsense mutations, which code for a stop and can truncate definitively the protein. Other small-scale mutations are insertions, in which we notice an addition of one or more nucleotides in DNA. Deletions, that are the absence of one or more

nucleotide from a DNA. From the other hand, we can distinguish those large-scale mutations in chromosomal structure that are involved larger regions (usually more than 1.5 kb). Analogous to small-scale mutations, we can distinguish gene duplications (or amplifications) that lead to multiple copies of those chromosomal regions, and deletions, so the removal of nucleotides, leads to a loss of the genes within those regions. Moreover, we also could detect the chromosomal translocations that are the interchange of genetic parts among nonhomologous chromosomes. Interstitial deletions that is a intra-chromosomal deletion that removes a segment of DNA from a single chromosome. Finally chromosomal inversions shows a reversal orientation of a chromosomal segment. Of course, all of this changes, are respect a reference genome.

Summarizing, when a point mutation occurs, a protein change, according to New England Review on genetics (2007), we could end up with this situations:

- No protein made.
- Too much or too little protein made.
- Misfolded protein made.
- Altered active site or other critical region.
- Incorrectly modified protein.
- Incorrectly localized protein (build-up of protein).
- Incorrectly assembled protein.

In all results, it is not expected than the system works as properly. In some cases, maybe a compensation system will found some paths to deal with the situation, and in others, not. Therefore, we can consider that a mutation, can result in disease if the mutation results in failure of the protein to correctly function. In order to assess if some mutation, is novo or rare mutation, or an SNP, it is necessary to keep track of all this kind of variants. Most of the times, you could predict the deleterious effect of the mutation if it is nonsynonymous, that is, the resulting codon codifies for a different aminoacid, or because it results to be a stop codon mark. Therefore, if the mutation changes the reading pattern or the mutation adds a stop codon, or the final codon codifies for an aminoacid with different physicochemical properties, we can argue that there could be a problem there. Nevertheless, if we have synonymous mutations, that is although the resulting different codon codifies for the same aminoacid, not always has a neutral effect. The reason is that certain codons are translated more efficiently than others. A

study with fruit fly alcohol dehydrogenase gene were introduced, changing several codons to synonyms, the production of encoded enzyme was reduced and the adult flies showed lower ethanol tolerance (Carlini et al (2003)). Another reason why synonymous changes are not always neutral is the fact that exon sequences close to exon-intron borders works as RNA splicing signals. But sometimes, regarding to have a synonymous mutation, it changes the original set of nucleotides, provoking deleterious effect on splicing, and could end up with an incorrect protein. This should end up with a truncated mutation that maybe not works at should it be (Pagani, F. et al (2008)). In that study found that about a quarter of synonymous variation affecting exon 12 of the cystic fibrosis transmembrane conductance regulator gene result in that exon being skipped.

A usual strategy in several cases is to predict if the resultant mutation could be deleterious and if it is, perform a pedigree study confirming that this particular mutation is not found on unaffected family members it is likely that this mutation leads to the disease. In the case of CNVs, this strategy could also be applied.

CNV

Copy-number variations (CNVs) are alterations of DNA that results in the cell having a different number of copies of one or more sections of the DNA. This provokes a clear imbalance of different genes, for a certain region.
For example, humans, as a diploid organisms we have 46 chromosomes, which consist of two sets of 23 distinct chromosomes. Because each individual possess a double set of chromosomes, it also has a double set of each of its genes. Any gene is located at a particular place on a chromosome (genetic locus). Therefore, in a given locus, we could have a set of genes coming from father, and in the same locus, a similar set of genes coming from the mother. In an hypothetical example, If we imagine that the CNV came from father's chromosome and affect that particular locus, and it was a duplication CNV event, we expect that in that individual, the expression of that kind of genes were higher than other individual with a single copy in each chromosome. If instead a duplication is a deletion, we'll end up with fewer dosage compared than normal. Note also, that if both chromosomes have a deletion it will provoke that certain genes are not produced anymore and usually it have even dramatic consequences by the organism. So it is known that differences in gene dosage will end up usually with a disease.

Besides that it could seem alarming to have this kind of imbalance in genome, most of CNVs are stable and heritable, so CNV between individuals is largely a product of genetic heritage. Indeed, deletions and duplications of chromosomal segments are a major source of variation between human samples. Nevertheless, the CNV could even be produced as de *novo* event at various stages of development. Multiple homologous recombination reactions are required for meiotic cell division that will end up with new haploid gametes, and although these events usually have high fidelity, occasional mistakes appear. This mechanism is known as non-allelic homologous recombination (NAHR) (Hurles, M. E. et al (2006)). Remarkably, not only in meiosis the CNV could be formed, other studies points out cases in which revealed extensive CNV between different cells in the same individuals; these CNVs must have arisen in post-fertilisation events (Piotrowski, A. et al (2008), Abyzov, A. et al (2012), McConnell, M. J., et al (2013)).

Other type of CNVs cannot be reconciled with non-allelic homologous recombination, and other mechanism has been proposed, such as rare replication defects resulting from broken DNA (Hastings, P. J., et al (2009)). One of those mechanisms is also called microhomology-mediated break induced replication (MMBIR), and seems to appear when cell is stressed and the repair process fail (Daley, J. M. et al (2005), Lieber, M. R. et al (2008)). In those cases could also appear translocations, that are regions of other chromosomes were copied. CNVs are not randomly distributed in the human genome and tends to be clustered in regions of complex genomic architecture. For example, in the regions which there are patterns of direct and inverted Segmental Duplications (or also called Low Copy Repeats (LCRs)), are more likely to be produced (Ross, M. T. et al (2005)). It is a particular case in which acts the already known mechanism of Non-allelic Homologous Recombination (NAHR). Other regions more prone to suffer these events are localized in telomeres and centromeres (Ross, M. T. et al (2005),Shao, L. et al (2008)).

Those kinds of imbalances usually end up with the development of certain disorders. A remarkable example is the Williams-Beuren syndrome (WBS), that is a neurodevelopmental disorder that is caused by a deletion of 26 genes from long arm of chromosome 7 (7q11.23) (Francke, U. (1999)). From this 26 genes, just to highlight

some of them : CLIP2, ELN, GTF2I, GTF2IRD1, and LIMK1. WBS is characterized by mental retardation, a learning disability, visuospatial cognitive problems and presents a highly social personality. Moreover, as several genes are affected is usual also to observe other phenotypic features in these patients, among characteristic facial features, usually present comorbidities such as full cheeks, coarse voice, hernias that may be explained by the insufficient supply elastin (ELN) (Jurado, L. A. P. et al (1998)). Strikingly, in the dimension of sociability, seems to be the opposite of autism disorder, but WBS patients have an impairment in cognitive function related with visuospatial functioning. The observation of opposite effect on sociability could be explained by those autism-like individuals, that precisely have a duplication in 7q11 (dup7q11) (Sanders, S. J. et al (2011)).

Mosaic events

Mosaicism is defined as "the coexistence of cells with different genetic composition within an individual, caused by postzygotic somatic mutation" (Rodríguez-Santiago, B. et al (2010)). That is, for a given sample, it is expected that for a given tissue, all the cells that build this tissue have exactly the same genetic code. In the case of mosaicism, this is precisely not the case.

In Fig 3.3 we can see the different scenarios that can lead to mosaicism (Campbell, I. M. et al (2015)).

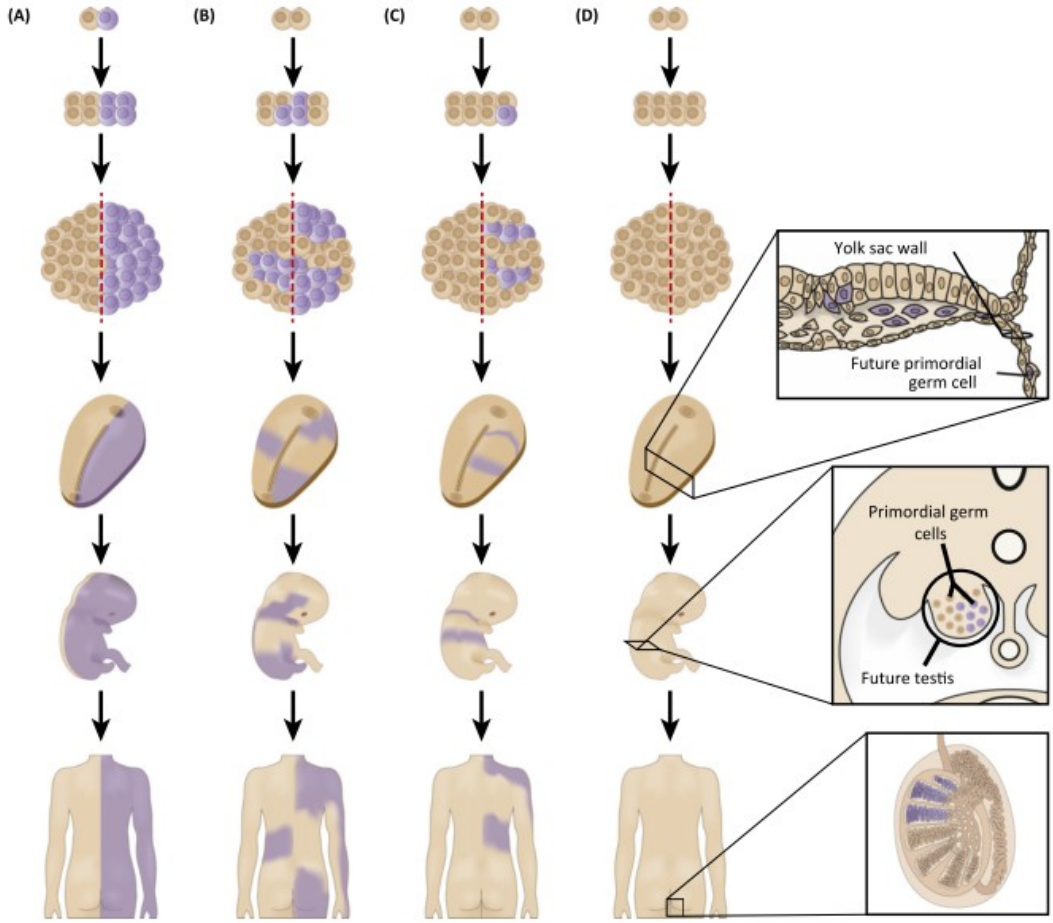


Fig 3.3 Timing of postzygotic mutation influences and distribution of mutant cells. A) Mutations that occur during the first mitosis. B) Mutations that occur before left–right determination can affect both sides of the individual, including one or both gonads. C) Mutations that arise after the determination of the two sides of the embryo can be confined to only one side of the individual. Only one gonad is likely to be affected. D) Mutations that occur after differentiation of primordial germ cells (PGCs) will be absent from somatic tissues. Scheme from Campbell, I. M. et al (2015)

Basically can differentiate between somatic (from Greek *soma* means body) mutations that will occur in any of the cells derived from the zygote (or the descendants of it) and conforms the different tissues and organs of the multicellular organism. Or those mosaic alterations that could happen in the germline mutations that happen in the gametes (or germ cells : sperm and ova). In this case, this mutation could be transmitted to the offspring.

In some sense, cancer could be viewed as a mosaic alteration (see Jacobs, K. B. et al (2012)), that is the cancerous cells the genetic content is different from healthy cells. In fact, the cancer cells present a lot of different alterations (large mutations,

translocations, etc.), and also is normal to detect in them mosaic mutations. So one might think in those cases, if mosaicism is the cause (driving force) or consequence that leads to cancer?

Despite this surprising feature of human genome, and in some tissues are a rare event, for example in young samples in blood, it is more frequent than expected in other tissues. Precisely it is normal that cells, when the immune system needs to create new antigens, usually some cells re-code some parts of their own DNA in the tissue where the infection is produced in order to produce antigens that recognize the pathogen. These changes are produced usually in specific place of the short arm of chr6 (HLA system). Other common forms of mosaicism are detected in karyotyping in pre- and perinatal embryos, and also, in single differentiated neurons, with average frequency of 1.25-1.45% per chromosome, with perhaps lower frequency on other cell types. When we talk about mosaicism events we highlight that occurs in the first stages of the development, when a mutation in a cell is propagated to a subset of an adult cell. The landscape of mosaicism in human tissues is still not so clear, and its frequency and extent in the adult normal population are still unknown.

The consequences of mosaic events are not so clear in most cases. Well known diseases that are likely derived by this event, is the case of monosomy of the X chromosome as a common type of mosaicism observed in normal individuals that are associated with aging (Guttenbach, M. et al (1995)). Other example is the recurrent lymphoblastic leukemia present in those patients with the monosomy of chromosome 7 (Russo, C. et al (1991)). Basically the consequences of mosaicism depend on the altered genetic architecture and how it affects in cell-specific pathways. Indeed, we have incomplete data of mosaic somatic alterations, as they can have no apparent phenotypic effect, and most of them can go undetected using the high-throughput genome analysis methods applied to DNA samples. As we already commented it is highly related with age, as more old is the sample, it is more likely to have more mosaicism events (in their own somatic cells, and also it is likely to have more mosaicism events in germline cells (Lin, G. N. et al (2015))). An enrichment in mosaicism events was detected in children with developmental disorders over healthy ones (odds ratio = 39.4 P-value 1.073e-6, 1303 over 5094 children lacking developmental disorders) (King, D. A. et al (2015)). Just to highlight that is a very uncommon event in children blood samples.

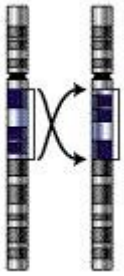
Thus, the relevance and frequency of mosaicism is likely underestimated in most of the cases, and we have to take into account.

b) Unearthed common variants

Copy neutral structure events - inversions (and submicroscopic inversions)

> "I believe in evidence. I believe in observation, measurement, and reasoning, confirmed by independent observers. I'll believe anything, no matter how wild and ridiculous, if there is evidence for it. The wilder and more ridiculous something is, however, the firmer and more solid the evidence will have to be.". Isaac Asimov

In the previous section, we see that the regions formed by a pair of near-identical sequences of low copy repeats (LCRs), could induce to produce CNVs. So these regions could be "hot spots" regions for unequal crossing-over lead to genomic mutations such as deletions, duplications or translocations. Besides those CNVs events, other chromosomal events could appear, such as inversions. An inversion is basically a chromosomal mutation present a change in a orientation of a given loci respect a reference orientation.



This figure shows an inversion, that is a chromosomal mutation in which a given loci change its orientation respect to a reference orientation.

Similarly to CNVs, this events is likely to be produced in zones with a enrichment of segmental duplication, although it can be produced by several mechanisms.

This event was discovered by a student of the famous geneticist Thomas Hunt Morgan and the creator of the first genetic map, Alfred Sturtevant . While he was studying the inheritance model on Drosophila, he realized that there are some genes that seems to be somehow related, thus they will be inherited together more frequently than expected. Moreover, this relation not depend on the physical distance, and coined a new measure : the centimorgan (cM). Centimorgan is a unit for measuring genetic linkage

between two chromosome positions, and one unit of centimorgan between this two positions represent the expected average number of chromosomal crossovers in a single generation was 0.01. Somehow could be used as a unit of distance, despite it does not rely in a physical distance, and in fact, he used to create the first genetic map (as previously commented, 1 cM is not a physical distance but at practical point of view it was found that 1 centimorgan correspond, on average, to 7.5e6 base pairs according to).

While he was building this map, he discovered that some of the flies have a different haplotype pattern than others (Sturtevant, A. H. (1921)) . Later on, Tan, C. C. (1935), confirmed the existence of the predicted Sturtevant inversions with microscope images. The chromosomes of the salivary glands of drosophila are formed with multiple rounds of replication of the sister chromatids that remain synapsed together. This fact produces those over-sized chromosomes (so called polytene chromosomes) let Tan to see, with the help of microscope, the inversions. An example of how inversions look like onf Fig 2.1.

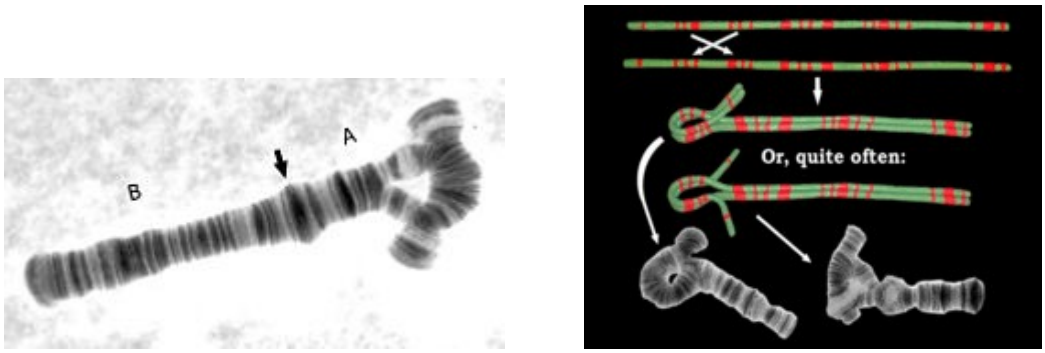


Fig 2.1 On first picture, between the point A and B, there is a bubble-kind regions, that in fact are the inversion. Next to it, a schema of how are they formed.

Alfred Sturtevant also observed that if they were heterozygous, inversions suppress recombination. This fact, can increase the risk of infertility. Therefore, inversions can module the recombination rate across the chromosomes.

Apart from the size of the inversion, inversions could be classified as pericentric (*peri-* comes from Greek language and means near of), that includes a centromere, or paracentric (*para-* means to place aside) that not. With pericentric inversions, a single crossover event that occurs between the breakpoints of a heterozygote produces unbalanced gametes that carry deletions, insertions, and either zero or two centromeres. This fact reduces fertility, making the inversions underdominant (lowered heterozygote fitness) as suggest the study realized by Coyne J.A et al (1991). The observation of a fixation of a given inversion among related species, lead to the idea that genetic drift can cause chromosome evolution in opposition to natural selection. Indeed, chromosomal inversions are found as polymorphisms within species or fixed in many groups of animals and plants. Therefore, an important feature of inversions lies in their ability to produce genetic isolation between populations and species.

Later on, some findings proved that some inversions can also be deleterious, such when the inversion breaks the usual lecture pattern of a given gene or critical regulators, or in the predisposition to cause rearrangements in offspring. These are the cases of germline rare inversion mutations, and some examples are the genomic traits as Hemophilia A syndrome (Russo, C. et al (1991)), Williams-Beuren syndrome (Osborne, L. R. et al (2009)), Prader Willi syndrome, or Hunter syndrome (Bondeson, M. L. et al (1995)).

But in most of the cases, the presence of an inversion seems to be neutral, such as those that appear at pericentric inversion of the chromosome 9 (9p11q13), or the inversion on 4p16 (Giglio, S. et al (2001)). The reason is that, despite this abrupt change over a region, the genetic material inside the inversion is balanced with no extra or missing DNA, and it could be translated normally. Despite in some cases heterozygous for an inversion individuals have lower fertility due to an increased production of abnormal chromatid, no more severe effects were expected.

Inversions have different sizes. Some of them, could be uncovered directly with the microscope (> 10 Mb of base pairs), while others, comprises from 2 nucleotides to several Mbs. Those inversions have less than 0.5 Mb could be detected by analyzing directly the genotyped sequence. In this phD thesis we are focused in studying, those ancestry submicroscopic size inversions (0.5Mb-5Mb) in human genome.

Formation mechanisms of submicroscopic inversions

Usually common submicroscopic inversions are produced via Non-allelic Homologous Recombination (NAHR), also briefly commented in previous section. But they also could be originated by double-strand break repair mechanisms, like non-homologous end joining, or replication based mechanism mediated with microhomology, like fork stalling and template switching (Lam, H. Y. K. et al . (2010), Kidd, J. M. et al (2010), Pang, A. W. C. et al (2013)). NAHR usually occurs between sequences of DNA that have been previously duplicated through evolution, and therefore have low copy repeats (LCRs). When non-allelic homologous recombination occurs between different LCRs, deletions or further duplications of the DNA can occur (Hurles, M. E. et al (2006)). We can see this mechanism in Fig 2.2 :

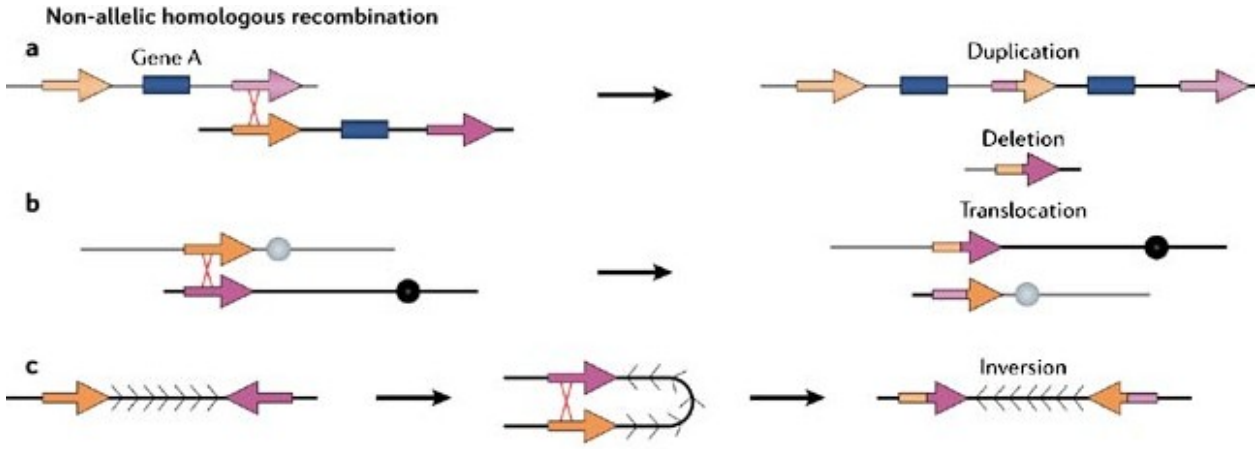


Fig 2.2 We could see a NAHR between two tandem SD (represented as arrows) in different possibilities. In **a**, we have that the result of NAHR between adjacent duplicated sequences are duplication or deletion event. Alternatively, translocations can result from exchange between SD on non-homologouse chromosomes as we can see in **b**. In both cases the SD orientation are the same. In contrast, in **c** situation, the segmental duplication are inverted and could produce inversions. (adapted from Bailey, J. A. and Eichler, E. E. (2006))

A reliable feature of human genome compared in different species, is that the frequencies of SD in human is higher. In different studies points that represent approximately 5% of human genome. These duplications are fragments with less of 1 kb

of DNA sequence and they can contain genes, pseudogenes and intergenic regions. Its origin is relatively recent (~ 40 million years), and have a high homology degree (~90%) (Bailey, J. A. et al (2002)). Besides that, we have to take into account that potential sequence misassignments with high identity levels of non-continuous clones may underestimate the real frequency of DS (Cheung, J. et al (2003)). The distribution of DS varies among the different chromosomes, as for example the Y chromosome contains DS among the 25% of its size and includes blocks from 1.45 Mb with a sequence identity of 99.97%. Moreover, subtelomeric and pericentromeric regions are rich in sparse SD as a result of translocations events. Taking into account this data, the predicted amount of common inversions in humans should be (small or large inversions) as many as around ~600-900 (Korbel, J. O. et al (2007)). Nevertheless few of them are really proved, due mainly as the difficulty in apply molecular biology techniques in regions that are rich in segmental duplications.

Main structural features of submicroscopic inversions

When heterozygous, inversions suppress recombination. Therefore, is only possible recombination events in regions that have the same orientation, otherwise the inversion mutation should lead to a non viable organism. This leads differentiated alleles between the inverted and the non-inverted allele.

In the Fig 2.3, it is shown schematically the process of formation of an inversion, and the evolution of it during N generations.

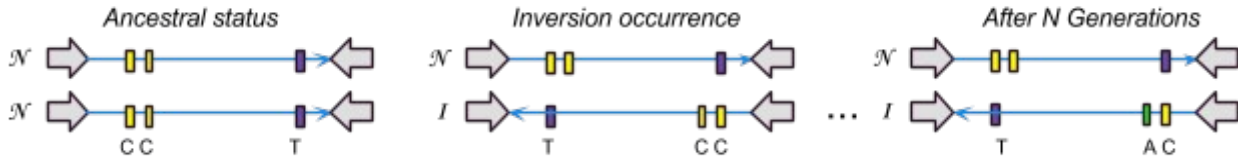


Fig 2.3. Different samples captured in different steps of time. As reference, some reference SNPs are colored in yellow and purple. First the inversion in ancestral status, before the inversion was produced. The next figure shows the formation of the inversion. Finally after N generations, this picture suggest that some new SNPs will appear in one chromosome but not in other.

Each inverted allele (N-noninverted, I-inverted), evolves differently during the time. Therefore, each allele tends to accumulate different kinds of mutations, and after multiple generations is possible to distinguish one chromosome with the inverted allele from another with the non-inverted. This fact makes submicroscopic ancient inversion as an excellent fossil biomarker.

Precisely, this feature lead us to apply computer methods to detect structural variants only looking for the SNPs that appear in a region in which we know that there is an inversion. More details are commented in *Methods* section.

Effect of inversions (and submicroscopic inversions as well)

Supposing that the genetic material was balanced, I want to highlight the main effects detected of the inversions :

Risk in meiosis : When heterozygous, inversions suppress recombination and causes subfertility (or directly infertility). Therefore this could contribute to speciation and reproductive isolation.

Variation selection : Each inversion allele has its own features, therefore at ecological level, could have an impact. An illustrative example could be the adaptation to aridity of *Anopheles gambiae* (Kirkpatrick, M. et al (2006), Simard, F. et al (2009)). Examples in human population are the case of the inverted allele of 17q21.31 inversion is under selection in european population (Stefansson, H. et al (2005)).

Broken reading frame : Inversion could have deleterious effect if a reading frame of some gene (or enhancer) was broken up. As in the case of Hemophilia A, that it is triggered in deleterious mutation over F8 gene on Xq28 (coagulation factor VIII). This gene presents a 9.5kb repeat in intron 22 that could provoke the apparition of an inversion. As a result the usual reading frame is altered, and appear that disease (Turner, D.J. et al (2006)).

Gene expression: As many studies pointed out, the inversion correlates with the expression of certain genes (Bosch, N. et al (2010), González, J. R. et al (2014)), and it is likely that have an impact on a given disease.

All those features are on review and in fact we are exploring some of them. Most of relevant works are found in ecological studies between similar species (or within same specie). The fact that inversions have this strong fossil mark, make them very appreciated between evolution researchers. Several studies are performed on certain easy-to-track huge inversions among several species such as *Anopheles Gambiae*, *Drosophila*, that provides relatively easy way to genotype them. Certainly, the fact that some of them (submicroscopic ones) are not so easy to detect, are retaining them for the moment, more literature on it.

Other feature to explore more in detail is the fact that present some risk in meiosis events. This suggest that they could have an impact in mitosis as well, generating mosaicism events among the affected cells. Could this variation explain some somatic alterations such as cancer?

It is known that all those mutation events, could cause dramatic effects on the organisms.
For example, well-known example of large-scale duplication mutation, is the as Down syndrome (Jacobs et al, (1959)), involves an extra copy (or sometimes partial) of the chromosome 21. Is typically associated with physical growth delays, characteristic facial features, and mild to moderate intellectual disability. Moreover, the affected with this disease have a high probability to suffer some kinds of diseases, particularly in heart, digestive system, and endocrine system. This is explained by the fact that there is an excess of synthesized proteins due to the extra chromosome material.

An example of large-scale deletion mutation is the Williams-Beuren syndrome (WBS), that is a rare developmental disorder (1-7,500-20,000 births) characterized by a deletion of about 26 genes at long arm of chromosome 7 (Martens, M. A. et al, (2008)). The effect of this deletion characterize a distinctive facial appearance, along with a low nasal bridge an unusually cheerful and ease with strangers. Additionally, the lack of this genes increase cardiovascular problem, such as supravalvular aortic stenosis and transient high

blood calcium. Regarding the sociability dimension, that is an excessively social interaction for the Williams-Beuren affected patients (WBS), there are examples in literature, than we could observe exactly the opposite effect, such as diseases that involve the autism spectrum disorder. Therefore, It is not a surprise that in recents years, several genetic evidences confirms this perception. An illustrative cases describe that a duplication on 7q11.21 on ADHD and autism affected samples could explain exactly the opposite behaviour (Somerville, M. J. et al (2005), Berg, J. S. et al (2007), Van der Aa, N. et al (2009)).

That examples suggest that different dosage alteration of 7q11.21 genes (LIMK1, GTF2I, GTF2IRD1 among others) have an influence on human language and visuospatial capabilities, among other imbalances. Therefore, the differential expressed genes could have an important role on some diseases.

c) Other variants

We tend to focus on things that we can easily measure. The genomics field is no exception, we tend to pay more attention in those parts that are already well-known annotated, or those that is easy to check, such as protein-coding regions (are the "functional" elements, right?!). But there are, at least, two other variants that we have to pay more much attention, and indeed, could have an important role with the disease.

One variant are those elements that are not protein-coding sequences, and fall in the so-called noncoding DNA regions. Historically the noncoding regions were catalogued as "junk DNA" (Susumu Ohno coined this term 1960) and were thought that most of the code in human genome have no functional effect (Palazzo et al (2014)). But recently, due the efforts to identify functional regions in human genome, and mainly by the Encyclopedia of DNA Elements (ENCODE project), suggest that 80% of human genome could be assigned to human genome. Nevertheless, still the proposal that in many eukaryotic genomes lacks an organism-level function (including the human, of course) and the concept that there is still a proportion of "junk DNA" (Ohno, S. (1972)) remains on the table (and still is provoking a lot of debate).

The other variant that seems to play an important role in (almost?)[2] all life genomes are those regions that are susceptible to be methylated or not (and therefore silenced or not). Indeed the epigenetic marks along the DNA vary among one individual, and still is continuously under developmental (either by self-reprogramming of DNA, either by environmental factors).

There is a growing interest in recent years in these two features of genomic landscape. I briefly comment those two alternatives in the following sections.

[2] By mass spectrometry revealed 14% of cytosines methylated in Arabidopsis thaliana, 8% in Mus musculus, 2.3% in Escherichia coli, 0.03% in Drosophila, and *virtually* none (< 0.0002%) in yeast species.
(Capuano, F., Kok, R., Blom, H. J., & Ralser, M. (2014)).

Repetitive elements, transposons, paralogous sequence variants

A large fraction of the genome consists of highly repetitive DNA. These regions are extremely variable even among the same population, and they could be used to identify uniquely a single sample. A main feature of those regions is that can be expanded or contracted through crossing over or replication slippage processes. As tandem repeats are unstable regions of the genome where frequent insertion and deletion of nucleotides can take place, could at the end, result in genome rearrangement or slippage (Viguera, E. et al (2001)).

Those repeats could arise also from the so-called transposable elements (TEs). TE is a DNA sequence that can change its position within the genome, sometimes creating or reversing mutation and altering the cell genome size. Barbara McClintock's discover these "jumping genes" that gives here a Nobel prize in 1983. TEs correspond to a large fraction of the C-value (or the DNA contained within a haploid nucleus) of eukaryotic cells, and before Barbara McClintock's finding, and regarding are generally considered non-coding DNA, it has been shown that TEs are important in genome function and evolution (Bucher, E. (2012)) and also play a role in gene regulation. The role of highly repetitive elements are not so clear, and currently there is no evidence that the majority of highly repetitive elements were functional (Cowley, M. & Oakey, R. J. (2013)).

Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding ability (Elio F. Vanin (1985)). Usually are mRNA of transcribed genes (so, without introns) that are inserted in DNA. Although not protein-coding, the DNA of pseudogenes may be functional (Poliseno L et al. (2010)), similar to other kinds of non-coding DNA which can have a regulatory role.

Altogether, those elements conform are commonly known as "non-coding regions" of DNA. Non-coding regions are those components of organism's DNA that they do not encode for protein sequences. During last decade, important efforts have been made in order to elucidating the roles of these "genes" in normal physiology and disease. Its proportion varies among the organisms, as it was illustrated in the following picture.

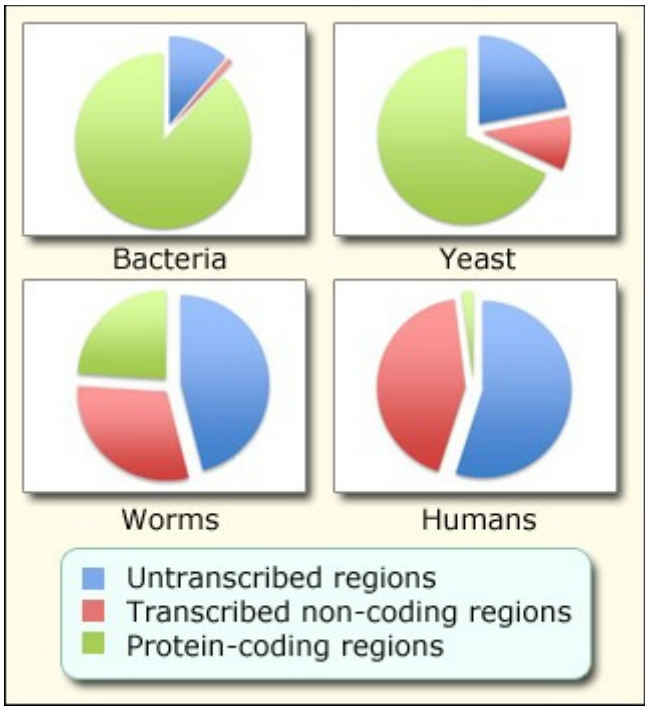


Fig 2.4 It shows the relative proportion of protein-coding, untranscribed and transcribed regions of four genomes. With permission of Nadya Dimitrova.

As we can observe in Fig 2.4., in differentiated organisms, we see an impaired proportion of transcribed, untranscribed regions and protein-coding regions. For example, over 98% of the human genome is noncoding DNA (Elgar, G. & Vavouri, T. (2008)), while only about 20% or less of a typical prokaryote genome is noncoding DNA (Costa, F. F. (2012)). This variation suggested that genomes can contain a substantial fraction of DNA in order to regulate genes.

ENCODE findings suggest that 80% of human genome have a biochemical function, based on their findings on 1% of the genome fully analyzed (Ecker, J. R. (2012)). It is not so clear if the non-coding parts of the genome have or not functionality, but according to Alexander Palazzo, and he claims that most of them, likely have no function based in some philosophical questions (f.ex "Onion Test") and some evidence (e.g. from many of them, "C-value paradox": the human genome seems too large, given the observed human mutation rate, and suggest that if the entire human genome were functional (in the sense of being under selective pressure), there will be too many deleterious mutations per generation). This observation, together with the fact that there

is a fail of correlation with genomes of different species, and the concept of "junk DNA" commented before, certainly gives a truly puzzling fact for scientists, and guarantees a lot of entertainment in the following years.

According to the Encyclopedia of DNA Elements (ENCODE), we can distinguish between long-range regulatory elements (such as enhancers, repressors, silencers, and insulators), promoters and transcripts. Therefore, besides protein-coding sequences, there are other that are transcribed into a functional non-coding RNA molecules such as tRNA, rRNA, snRNA, lncRNA (e.g. ENCODE project, such as Birney, E. et al (2007)). Are thought to be regulatory regions that control gene expression together with promoters, silencers, and enhancers. It is estimated that human genome have about 12,600-19,700 genes (Alexander F. Palazzo (2014)).

Indeed there are strong evidence that non-coding RNAs (ncRNA) play an active role into some diseases as pointed out in the review of Esteller, M. (2011), and therefore it certainly we have to take into account.

Epigenetic

In its origin, epigenetics was used to describe events that could not be explained by genetic principles (Conrad Waddington 1905-1975). The epigenetics (epi- in Greek means over, outside of, around) is based on the trait variations that are not caused by changes in DNA sequence, but those external or environmental factors that turn genes on and off and affect how cells read genes (Holliday, R (2014)). In recent years, they show that the genome was more dynamic than thought. It was already known that there are portions of DNA genome that are physically changing (e.g. immunological HLA region in chr6), but the living-organism have other system in order to silence or not certain genes of a given region without the modification of genetic code.

There are different epigenetic mechanisms that cells use to control gene expression. DNA methylation is one of them (not to be confused with histone methylation), and basically is attachment or substitution of a methyl group in a specific region, provoking the silencing or not of a given gene. Methylation is important in numerous cellular processes, including embryonic development, genomic imprinting, X-chromosome

inactivation, and preservation of chromosome stability. The core concept is that the effect of have a cytosine unmethylated is that it creates a permissive chromatin environment, whereas the other way around, if it is methylated not (Blackledge and Klose (2011), Deaton and Bird (2011), (Morris, K. V. (2014))). Therefore, it can operate from a single gene to hundreds of them.

Others epigenetic mechanism involves chromatin and histone modification, with regulatory RNAs.
Those epigenetic marks could also be inherited, and are something dynamic during the life of an organism. Several studies shows that epigenetic marks changes during the life of the persons, and also that correlates their number among the ages (Bell, J. T. (2012)).

The principal actors in the control of methylation events are thought to be the ncRNA. The mechanism of actuation could be different, based on the type of ncRNA. The classification historically was performed based on the size, and just to mention some of them(He, L. & Hannon, G. J (2004), Esteller, M. (2011), ENCODE):

miRNA: (small, ~22 nucleotides). Mediates gene silencing by controlling translation of mRNA into proteins. They are involved in regulating many processes, including proliferation, differentiation, apoptosis and development. Some miRNA regulates specific targets, while others could regulate the expression of hundreds of genes simultaneously. The known mechanisms to repress mRNA via miRNA are by mRNA degradation or by inhibition of translation initiation .

piRNAs: (Piwi-interacting RNA) : (small 24-40 nucleotides). Forms RNA-protein complexes through interaction with piwi proteins, and are involved in maintaining genome stability in germline cells. They have been linked to both epigenetic and post-transcriptional gene silencing of retrotransposons and other genetic elements in germline cells.

snoRNAs: (intermediate-size, 60-300 nucleotides) Are small nucleolar RNAs predominantly located in introns. They play a role in alternative splicing of the pre-mRNA in order to create the mRNA. After the splicing process, they could

be exported, working as a ribosomal RNA (rRNA), and it is thought that also could develop other functions (still unknown).

lncRNAs : (large, > 200 nucleotides) Are involved in many biological process, and this class of ncRNA makes up the largest portion of the mammalian non-coding transcriptome. They are known to mediate epigenetic modifications of DNA by recruiting chromatin remodelling complexes to specific loci.

All those ncRNA, and many others, have an active role in genome regulation. Given the critical role of DNA methylation in gene expression and cell differentiation, it is thought many process in which methylation plays a part, errors in this process could drive to devastating consequences, including several human diseases.

2. HYPOTHESIS

A significant proportion of the genetic variability responsible for multifactorial disorders and traits may lie in complex regions of the genome that have not been well analyzed with the methods available for genotyping because of their complexity (or because still we are not be able to detect them). We can improve tools and methodologies in order to detect and characterize better inversion chromosomal alteration using SNP arrays. There exist some tools to detect but we can add biological information in our algorithms to be more accurate.

We believe that a portion of missing heritability could be found is those under-unexplored chromosomal variations such as inversions. We believe that their study might provide us relevant knowledge about the molecular basis of multiple human disorders.

3. OBJECTIVES

The main objective of this PhD is to contribute to the missing heritability of human traits and multifactorial disorders by the analysis of several poorly explored sources of genome variation.

The specific goals are:

1) Optimize methods and tools to detect genotype and characterize ancient submicroscopic inversions of the human genome. This include localize the putative breakpoints, and define correctly the haplotypes, and tag them with available SNPs.

2) Study the inversions exploding the already available genomic data (dbGAP and other sources). Correlate the inversion with gene expression in available datasets in different tissues.

3) Define whether any of the detectable inversion could have a role in neurocognitive and autoimmune complex disorders. Characterization of submicroscopic ancestral inversions and their functional consequences and their putative implication in neurodevelopmental disorders.

4) Analyze other structural variants (copy number and copy neutral) posibly present in mosaicism as additional sources of genetic variation associated to the same complex disorders.

4. METHODS

> "Constructing better instruments is how the science mainly progresses. Measuring". Edwin Hubble

SNP arrays: squeezing new juice from old lemons

In 1975, Frederick Sanger develop a sequencing DNA method, commonly known as Sanger Method. The input is an isolated DNA chain that you wish to know their nucleotide sequence (you could use restriction enzymes to select it). Then, several reactive are added to primer and DNA template (with the complementary primer) and a little quantity of the four dideoxinucleotides (ddNTPs) marked with flurophores (different colour for each nucleotide: G, A, T, C), and other nucleotides. The main property is that, if some of those ddNTP flurophores is choosen by transcription factor, the DNA polymerase stops the transcription of the chain. The process is repeated several times, and it the end, by randomness, you end up with all the different possibilities, with a lot of complementary chains of the basic template, with a different sizes. And at the end of each chain, with normal nucleotides, and with a ddNTPs marked on a different colour based on the kind of nucleotide that it was in their tail (3'). The next step was simply sort (and cluster) those final chains based on length (using for example single lane capillary gel), and read the flurophores in that particular order. In this way, you can deduce, for every particular position the kind of nucleotide that it was, and therefore genotype the original chain (taking into account that you get the complementary). In Figure 4.1 there is a Sanger method schema. It is a nice example on how to use the well-known DNA mechanism that uses to be replicated, and only using restriction enzymes, primers, and core DNA machinery (DNA polymerase) and a pool of nucleotides, you could end-up with the original nucleotide sequence.

A restriction of Sanger sequencing is that only works for ~700 bp. With larger sequences is technically difficult, but fits well to check small ones.

From those times to nowadays, Sanger sequencing evolved and different strategies, based on the same core concepts (hybridization, restriction enzymes, replication, etc.), and with the desire to apply it to large-scales, appeared the so-called Next Generation Sequencing methods.

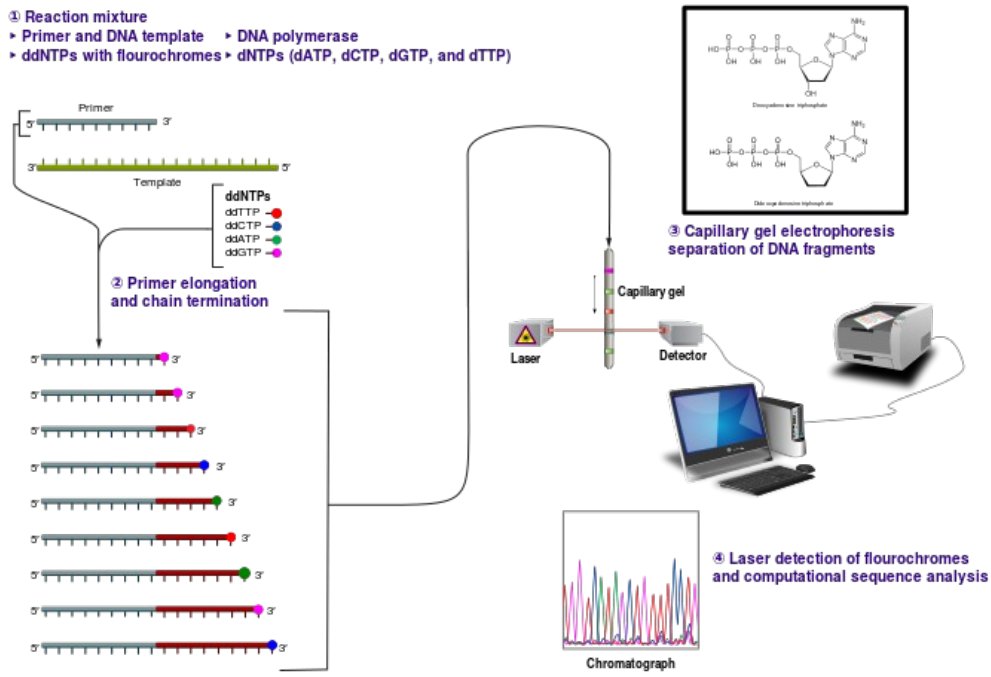


Fig 4.1 This schema shows the sanger method. Starting for the template with the primer that you want to deduce, and start reaction with DNA polymerase, and the different ddNTP with the different flourochomes, such as dATP, dCTP, dGTP and dTTP. When the process arrives to the end, a final chains with different termination are obtained. Those were sorted based on their length, and readed sequentially, generating a chromatograph. It is read, and it is possible to deduce the original sequence (source wikipedia:Sanger Method).

This provoked a revolution in genetics and human medicine, as it was possible to got the first fully sequenced human genome. Moreover, it is possible to sequence whole human genome relatively fast, in question of days. This high-throughout technologies, not only focused on DNA sequencing, but using exactly the same principles, was also possible sequence RNA (the transcript data -RNASeq) and also other DNA features, such as methylated regions. Lots of findings arose from collected data, lot of them also shaking. For example the similarity between human specimens, in which most of the common variants between human samples resides on ~0.1% (1000 genomes consortium) of the full human genome (about three billion bases). Besides that, the difference between humans, and their closest living relative (common ancestor was about 5-7 million years (Mya) ago) , chimpanzees accounts for ~4% (Ajit Varki and Tasha K. Altheide (2005)). The harvested data gives us a lot of information between and within species of different living organisms and shaped the strategies performed.

Precisely, from those initial findings, motivates, in the biomedical field, the creation of technology capable to study those chromosomal variation, and doing it al large-scale. As commented before, the vast majority (~99%) of genomic sites looks exactly the same among human DNA samples. Indeed, it motivates the collection of those SNPs in a single database), and appeared the dbSNP database of NCBI (Sherry, S. T. et al (2001)). Therefore dbSNP collects those SNPs that have a minimum frequency among population of 1 % (at 2014 were about 112,736,879 SNPs dbSNP build 132). And as many of them are in Linkage Disequilibrium between them, it can be reduced to 1-2 Million SNPs (those SNPs usually fall in noncoding regions, although also there are SNPs in some coding regions -exons).

In the last decade until nowadays, the central aim in human genetics was focused in study those DNA variants that could contribute to disease. Discover these causal loci requires the ability to assess DNA sequence variation on genome-wide scale. This fact, motivated to detect, as cheaper as possible, the point variation among samples, and appeared the Single nucleotide polymorphism (SNP) microarrays (SNP arrays). SNP arrays let us to genotype human DNA at thousands of SNPs across the genome simultaneously. Moreover, since their initial development, the platform's applications have expanded also to include the detection and characterization of copy number variation (whether somatic, inherited or de novo) as well as loss-of-heterozygosity. The availability of use genotyped data from healthy and unhealthy samples provide the opportunity to deeply study their phenotypes. That procedure provided (and still provides) new knowledge that led us to sketch some blueprints to understand the underlying architecture of the disease. The main motivation is design new strategies at molecular level that hopefully help to treat some of them. This strategy turns to be successfully regarding common SNPs variants directly implicated to some biochemical function (f.eg.rs148649884 and rs138055828 in serum levels of M-ficolin, Ammitzbøll, C. G. (2012)).

Of course seems inevitable that in a few years, the DNA arrays were displaced by Next Generation (NGS) technologies, and several studies points out the limitations of SNP microarrays and strengths of NGS data (Thomas LaFramboise (2009)). The Next-generation sequencers available (Illumina/Solexa, Roche/454 and ABI, with other on

the way) are able to produce all of the information that SNP arrays can produce, but with (theoretically) greater resolution and accuracy (Shendure (2008)). But still today, the maturation of SNP array seems a reliable option in many situations (for example, is possible to design on demand panels to look to check specific feature among samples).

But lot of data is already generated from SNP microarrays and it is freely available today. Indeed, SNP microarray is a mature technology and lot of effort and new computational methodologies developed to handle the resulting data. Indeed novel computational methodologies are still under development, providing the opportunity to exploding unexpected possibilities. For example, lots of GWAS performed the typical case-control association studies over single SNPs. But from raw data generated from SNP array is also possible detect rare variants such as CNV, Identity-by-descent regions and mosaicism events. Those possibilities are fully explored in the following sections.

In the following sections I will comment the hidden possibilities that microarray data could reveal such as CNV, mosaic event, and specially, as the main phD thesis title suggest, in the detection of submicroscopic ancient common inversion variants.

Inversion detection

The usual way to detect inversions with biomolecular cytogenetic technique is known as *Fluorescence In Situ Hybridization* (FISH). The fact is that you design primers with fluorescent probes that bind in specific sites of a given chromosome. .Then, with the electronic microscope is possible to view directly the colored pattern of those cells in metaphase to try to predict the inversion genotype.

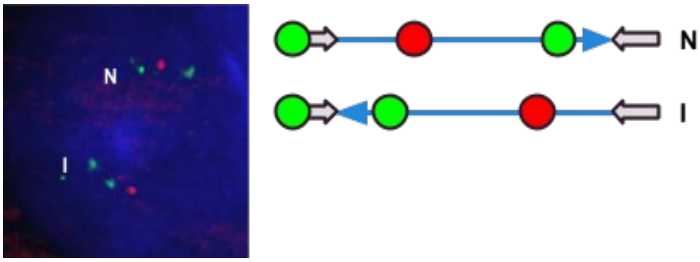


Fig 4.2 The two chromosomes of a single cell, that appear . (courtesy of Judith Reina).

In this way, you can infer the genotype of the inversion, that is inverted or not, based on the colored pattern.The drawback of this system is that due that the chromatin is not always sticked, and also the photo is bidimensional (and the reality is obviously in 3D) sometimes it is not clear than a given pattern correspond to a given orientation. Therefore several nucleus must be checked in order to genotype effectively that particular sample. Besides that, sometimes, due the landscape of the region, could be some problems to design primers, due for the repeats (segmental duplications) or for the genetic particularities of the zone. Also, for little inversions (less than 0.5 Mb) is not effective as the spots are very near one from each other. But in the most favorable scenarios, it is hard and time-cost effective method. This fact, motivates to search for high throughput methods in order to try to detect them.

Despite the success in paired-end methods in sequencing the full human genome in this last decade, somehow it is not enough in identify properly the inversions. A first limitation of that approach it that it relies on the reference assembly, so it is constructed with a mix of several samples. So the reference assembly it represents very rare or unique alleles at some loci in the genome, and also in some cases these unique alleles could be cloning artifacts or mis-assembly of the reference sequence. Also it is not possible to capture all variants in a single reference assembly. For example, for regions where the reference assembly harbors a unique inversion allele, a single homozygous inversion will be detected, independently of the amount of resolution and sequence coverage.

Other limitation is related with the inner architecture of the microscopic inversions, is that as commented, are usually flanked by high identity segmental duplications (> 1kb), that exist in two or more copies of more than 90% of identity in the human genome. As the method depends on the alignment to the reference assembly, those sequences will cause problems in identifying unique placements for sequence reads. Many paired-end mapping pipelines simply discard those reads that cannot be uniquely mapped. Therefore, the paired-end mapping strategy often fails to identify inversions flanked by long inverted segmental duplications.

And regarding other limitations of the technology, for example, sequencing by synthesis (Illumina) or by ligation (solid), or pyrosequencing usually provides 300 bp by reads (700 bp as many nowadays), so they are still too small reads for capture complete inversion sequences of 0.5Mb-4.5Mb. The newest technologies such are those provided by Pacific Bio, promises 10 Mb-15Mb for read, but for now are very expensive and still is under development. Therefore, one approach could be use the base read pair-end reads, and try to perform de-novo assembly per sample, per region, and try to detect those abrupt changes near the breakpoints. As it was the case in the detection of 17q21.31 breakpoints This make a hard task to find and correctly characterize the inversions. Actually, although several hundred submicroscopic inversions have been report in humans, only few of them (<15) have been characterized in greater detail, such as : 17q21.31 (Steinberg, K. M. et al (2012), Stefansson, H. et al (2005)), 8p23 (Salm, M. P. A. et al (2012)), and 16p11.2, 7q11.23 (Antonacci, F. et al (2009)).

Our approach is based on take advantages of well-known used technologies such as microarrays, and try to squeeze them as much as possible. A great advantage is that there are a lot of studies that are already performed with them, and for our study it have enough quality to, in most of the cases, predict and genotype inversions for the available samples.

The key point is that at least, we are able to detect those ancient submicroscopic inversions due those traces left on the microarrays. The fact that if a given inversion, is enough old, the inverted allele, compared to non-inverted allele, tends to accumulate different mutations. Therefore, during the time, are both more and more differentiated. So the algorithmic methods recently developed (Salm, M. P. A. et al (2012)), that relies on the detection of main haplotypes in a region based on multi dimensional scaling (MDS) and the detection on the differences on linkage of different SNPs close to the breakpoints of the inversion (Cáceres, A. et al (2012)).

We mainly used the same strategy, using two algorithm recently developed in order to detect submicroscopic ancient variants inversions that is *inveRsion (Cáceres, A. et al (2012))* and *InvClust* ((Cáceres, A. et al (2015)) an attempt to improve *PFIDO* algorithm (Salm, M. P. A. et al (2012))) developed by Alejandro Cáceres and Juan Ramon González.

InveRsion algorithm predicts abrupt changes in the pattern of LD between two contiguous genomic blocks flanked by a putative breakpoint. Once this region is susceptible to contain a submicroscopic ancestral inversion, with *invClust*, helps to identify the two main haplotypes of the inversion, and genotype each sample based on this classification. Therefore, if a region is predicted to likely to contain the presence of an inversion, and it looks like an inversion, it is likely to be an inversion. In following section there is a brief explanation of how these two tools work.

InveRsion

InveRsion algorithm scan genomes and predicts putative regions susceptible to contain inversion breakpoints. It mainly explodes the fact that the LD between the SNPs of two flanking genomic blocks could vary in some samples in contrast with others.

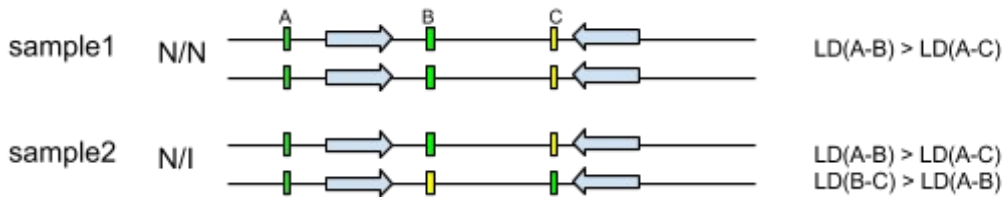


Fig 4.3 In this figure we see the schema of two different samples in a given region susceptible to be an inversion. The SNPs are labelled as A,B,C and mapped according the reference genome. Sample1 is genotyped as N/N, and therefore the SNPs A and B are in stronger LD than A and C, as expected based on its a priori proximity. In contrast, sample2 shows stronger LD between SNP A and C in one allele, then suggesting the individual is heterozygous for an inversion with a breakpoint between A and B. Segmental duplications are represented by light blue arrows.

In Fig 4.3 shows a particular example of two samples, in which the first sample is genotyped with N/N (Non-inverted allele/ Non-inverted allele) and the other one with N/I (Non-inverted allele/ Inverted allele). The second sample strikingly shows that SNPs near the B SNP are in LD with "far-away" SNPs , such as C SNP.

Therefore with a given amount of genotyped samples for a given inversion, *inveRsion* algorithm detects abrupt changes of LD between blocks. Concretely between SNPs of the two blocks. It is possible to scan a genome with different windows sizes in order to search for inversion signals without previous knowledge of the breakpoints. *InveRsion* algorithm quantify this positive signals using Bayes Information Criterion (BIC)

indicating that the chromosomes of some individuals between the breakpoints are more likely to be inverted than not.

It is an extremely powerful tool to detect those putative regions that likely contain inversion, but it demand a high processing resources and it have a high time-cost process. *InvClust* on the other hand, it is a lightest process in order to genotype samples, when you know the region that it is likely to be an inversion.

InvClust

InvClust uses a MDS in order to cluster the haplotypes that appear in certain region. The observation is that if we perform a MDS inside the inversion, the two haplotypes are detected.

It check that the selected SNPs were in HWE (Hardy-Weinberg Equilibrium) in order to establish that the genetic composition of a population remains in equilibrium (allele and genotype frequencies in a population will remain constant from generation to generation), in absence of other evolutionary influences. These influences include mate choice, mutation, selection, genetic drift, gene flow and meiotic drive (genome will effect a manipulation of a meiotic process in such a way as to favor the transmission of one or more alleles over another).

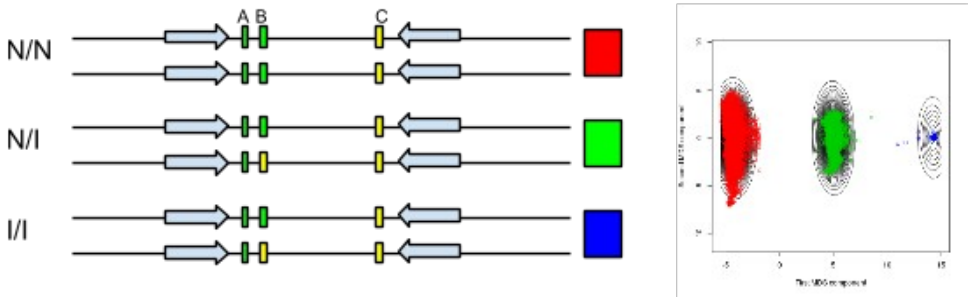


Fig 4.4 InvClust MDS clustering different samples according the similitud (distance) between the main haplotypes. In this figure, invClust detected 3 haplotypes (2 main clusters, so the inverted and the non-inverted allele) Segmental duplications are represented by an arrow.

Note that if a invClust is launched using SNPs outside the region flanked by the segmental duplications, producing this two-main inverted haplotypes (three clusters) as it shows Fig 4.4. If invClust is applied in a region that in fact there is not an inversion, the more likely outcome is to have a single cluster.

It is usual that in a site that there is an inversion, inverted and non-inverted alleles present different recombination rates (f.e. inversion 8p23 Joao M Alves et al (2014)). In a typical situation in which there is an ancestral inversion on a given region, performing a PCA/MDS with the available SNPs in the region, the probable outcome should be that there will be two clusters. In fact, InvClust always assume that by default there are two alleles (3 clusters). But there are some cases, such as in 16p11, that clearly appear three alleles (6 clusters). In those cases our hypothesis is that there should be an inversion inside other inversion.

In any case, when you have correctly defined the clusters, if there is available trios, is recommendable to verify that certainly that there is no violation of Mendelian inheritance in those patterns, either no HWE between unrelated individuals with the given putative inversion. In the cases in which there is more than 3 clusters, was try to define correctly the clusters (several methods are available, f.eg the R packages such as kmeans, Mclust, htclust in order to detect the different clusters).

Nevertheless, more work is required in order to demonstrate that there is a inversion on that zone. The second step was an experimental validation using molecular FISH or with the available BACs of the region with samples with inverted and non-inverted allele.

CNV detection

As commented in the structural variants section, CNVs are alterations that, in a given cell, we could have different number of copies of one or more sections of DNA. This variation comes with different flavours such as deletions, duplications, translocations and insertions. In this phD we only are focused with the detection of deletions and duplications regarding CNV alterations with the use of high resolution oligonucleotide array platforms.

The platforms used in the work performed in this phD are from Affymetrix, Illumina and aCGH arrays, with about ~1-2 Million of SNPs detected in human population. The strategy that those method follow are based on the intensity of the probes (*logRR*) and

allele frequency (*BAF*) that those SNP array provides. In order to perform this CNV analysis were use mainly *pennCNV* (Wang, K. et al (2007)), *gada* (Pique-Regi, R. et al (2010))and *parseCNV* (Glessner, J. T. et al (2013)).

All detection methods based on snp arrays mainly use the BAF and *LogRR* values. The *LogRR* is the amount of normalized detected signal of a given chromosome that hybridize within a given probe. This value usually is represented with *logRR* and usually is a value from -1.0 to 1.0, and informs the total copy number. In normal samples, in the major part of genome regions, those values should be around 0, informing that there is no signals of deleted or duplicated region.

On the other hand, BAF is derived from the ratio of the SNP signal intensities observed, and may be interpreted as the proportion of chromosomes carrying B allele. That is, for a single SNP, it captures, for normal samples, traces from paternal and maternal chromosome for the same region. In normal diploid sample, the expected BAF values are 0, 0.5 and 1 (corresponding to BB, AB and AA). f.eg. the paternal chromosome could contain the major allele A in that region, and the maternal chromosome the minor allele B.

Graphically we can distinguish between them in Fig 3.1.

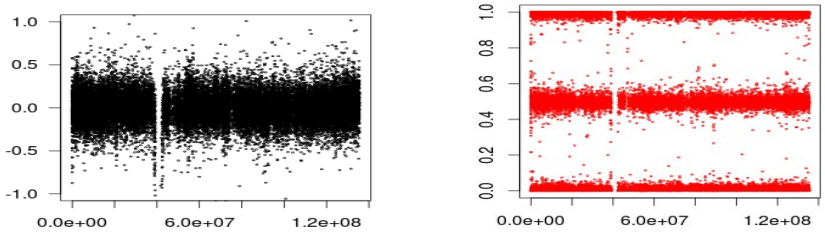


Fig 4.5 black a representation of *logRR* signal of all SNPs of a given human chromosome, and in red the *BAF* values.

In this way, those algorithms tries to identify if there is a gain (duplication) or a deletion based on *BAF* and *LogRR* values. For example, if in a given zone there is a loose on LogRR we can infer that there is a deletion or if we have a gain we can infer that there is a duplication. In figure 4.5 shows how *gada* applies this idea.

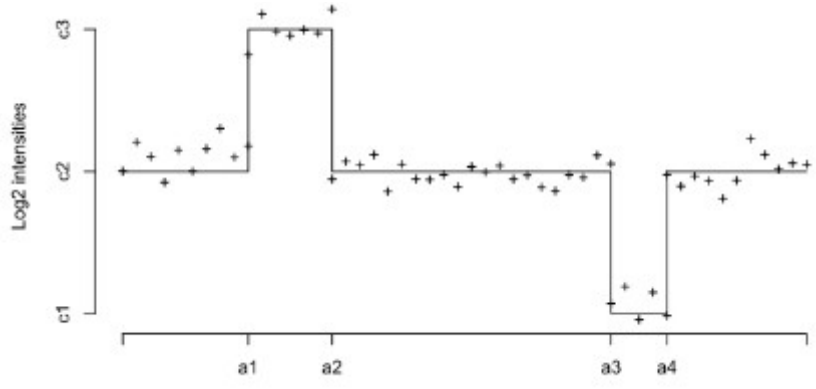


Fig 4.6 Diagram in which *y*-axis represents the log2 intensities, and in *x*-axis there is the positions of the different SNPs. Therefore in region [a1,a2] there is a duplication, and in region [a3,a4] a deletion.

For more details, a complete description on how *pennCNV* and *gada* works (Wang, K. et al (2007), Pique-Regi, R. et al (2010)).

Once the putative CNVs are found, other critical step after the routinary filters to discard those samples that are contaminated or that present too low genotype quality, it is critical to detect those CNV that could be artifacts. For example, those homologous regions could give calling errors. As an instance, if there is a zone with segmental duplication detected, it could be incorrectly assessed as a duplication when in fact this SNP region maps to some other place in the genome, and the DNA of the sample hybridize more in this point that in the other place. Usually this kind of artifacts are also detected in regions that are proximal to centromere and telomere. In order to detect those kind of artifacts and others, we used manual curation and the application *parseCNV* (Glessner, J. T. et al (2013)).

When the CNVs are detected it is also important distinguish between if they are common variations between the population or not (mainly in association studies). The main source of information used was extracted from the manually curated database of genomic variants (DGV) (MacDonald, J. R et al (2014)).

Finally, remark that CNVs (or structural variants) are more obviously difficult to detect than SNPs. SNPs always occur in two alleles, whereas there is about 5% of human genome defined as structurally variant in normal population. That represents ~800 independent genes, and indicates that structural variations can comprise millions of

nucleotides of heterogeneity within every genome, and are likely to make an important contribution to human diversity and disease susceptibility. Therefore, it is crucial to control the percentage of rarity (abnormality) of the variation, and make those analysis with many samples as possible.

Mosaic events detection

As commented in Mosaic section, mosaicism results from a postzygotic mutation, usually during development, that is appear in a subset of adult cells. It can occur in either or both somatic and germline cells, the latter with the potential of passage to offspring (Rodríguez-Santiago, B. et al (1999)).

We mainly used *mad* and *pennCNV* tools in order to detect mosaic events. In fact, the detection of mosaic events, follow a very similar strategy than CNV, as those algorithms uses *BAF* and *LogRR* of every SNPs values as well.

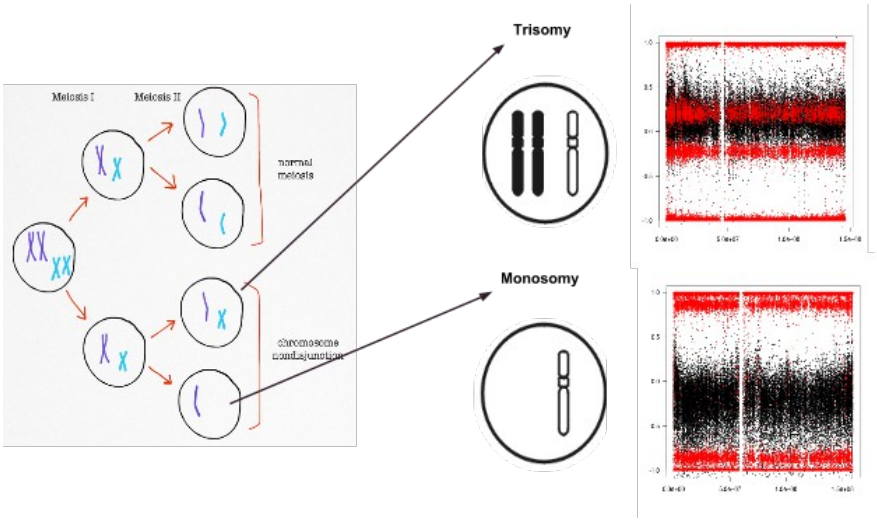


Fig 4.7 Two different types of mosaic events : trisomy and monosomy. *BAF* values are in red, and *LogRR* in black. Ideally the LogRR in both cases should remain centered.

The difference of CNV and Mosaic event is subtle, but evident as it shows on Fig 4.7. We see a pattern similar as CNV, but with the main difference that the *LogRR* is not deviated at top or bottom.

5. RESULTS

"You can get anywhere if you walk enough.". Lewis Carrol

5.1 Common inversion polymorphisms under selection are susceptibility factors for autism and schizophrenia

Armand Gutiérrez-Arumi, Alejandro Cáceres, Marcos López-Sánchez, Ivon Cuscó, Juan R. González, Luis A. Pérez-Jurado
Common inversion polymorphisms under selection are susceptibility factors for autism and schizophrenia.
(Submitted)

Armand Gutiérrez-Arumi[1,2,3], Alejandro Cáceres[2,4,5], Marcos López-Sánchez[1,2,4], Ivon Cuscó[1,2,3], Juan R. González[2,4,5], Luis A. Pérez-Jurado[1,2,3]

1Genetics Unit, Universitat Pompeu Fabra, Barcelona, Spain

2Hospital del Mar Research Institute (IMIM), Barcelona, Spain

3Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain

4Centre de Recerca en Epidemiologia Ambiental (CREAL), Barcelona, Spain

5Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

Abstract

Despite a major genetic contribution to their etiology based on twin studies and epidemiological data, a significant proportion of the heritability for autism and schizophrenia spectrum disorders (ASD and SSD), mainly attributable to common inherited variants, is still missing. We have genotyped the haplotypic signatures generated by suppressed recombination of four common candidate ancestral chromosomal submicroscopic inversions, using available data from several genome-wide association studies. The H2-allele at the 17q21.31 0.9Mb inversion was found over-transmitted to ASD probands of the Autism Genome Project trio dataset (11%, p=3.2e-04), mainly from fathers, in high functioning and verbal ASD and in multiplex families. In contrast, the same H2-haplotype was nominally associated with protection for SSD (OR=0.86, CI=0.78-0.96, p=0.028). Over-transmission of the I-allele at the 8p23.1 4.5Mb inversion was also observed in three independent ASD datasets (combined transmission 7.1%, p=2.5e-05). A milder transmission distortion of the same

alleles, significant for inv8p23.1 (combined transmission 5.8%, p=0.009), was also observed in normal siblings of ASD probands as well as in unrelated datasets of children with asthma and attention deficit hyperactivity disorder, suggesting the distortion is a general event caused by either meiotic drive or viability selection. Inversion alleles at both loci strongly associated with differential expression of several regional genes in specific brain areas and blood cells. Therefore, two common chromosomal inversions at 8p23.1 and 17q21.31, unevenly distributed among human populations due to genetic drift or adaptive selection, underlie part of the hidden genetic susceptibility to ASD and SSD, most likely through modulating the expression of regional genes.

During the last two decades, there has been a major drive and effort to identify the genetic variants that account for the acknowledged heritability of common phenotypic human traits and disorders. Single nucleotide polymorphisms (SNPs) are still the most studied genetic variants and the most common form of genetic variability in the population, along with structural variation affecting copy number[1-2]. However, they individually account for a relatively small amount of the heritability of complex diseases, even when additive small effects of multiple variants are taken into account[3-5].

Autism spectrum disorders (ASD) are neurodevelopmental disorders characterized by impaired social interaction and communication skills along with restricted and repetitive behaviors. ASD have a prevalence of ~1%, being more frequent in males than females. There is strong evidence for a genetic etiology with heritability ranging from 50 to 90%[6-7]. Schizophrenia spectrum disorders (SSD) are defined by abnormalities in one or more of the following five domains: delusions, hallucinations, disorganized thinking, disorganized or abnormal motor behavior, and negative symptoms. SSD share several phenotypic features with ASD, as well as similar prevalence and a high heritability estimate (50-70%)[8]. However, multiple genetic and genomic studies have revealed a huge heterogeneity in ASD and SSD through the identification of hundreds of genes implicated by single gene mutations, copy number changes and double hit models, explaining the etiology of roughly 1/4 of ASD cases[9], lower proportion in SSD. Although common inherited variation is estimated to contribute to between 40 to 60% of the variance[10-11], Genome-wide Association Studies (GWAS) have failed to identify common genetic variants consistently associated with ASD and SSD. Reasons for this poor success rate are the underlying genetic and phenotypic heterogeneity of these disorders and the possible underpowered of GWAS to detect small effects unless much larger sample sizes are used. Attempts to reduce phenotypic heterogeneity by subphenotyping had returned very low effect on genetic homogeneity[12].

We hypothesized that part of this missing heritability attributable to common genetic variation could be hidden in poorly explored and functionally relevant genomic variants. Chromosomal inversions, copy-neutral changes in the orientation of chromosomal segments with respect to the reference, are likely common sources of genetic variability[13]. Some common submicroscopic inversions have shown traces of selection in the general population[14-16], differences in gene expression, and some have

already been associated with multifactorial diseases[16-17], including susceptibility to the joint comorbidity of two complex disorders mediated by the concurrent effects on more than one gene by the inversion[15]. There is an updated database with 1092 reported inversions in humans, most predicted with paired-end mapping but only 85 validated, 20 larger than 0.2 Mb (InvFEST)[18]. Due to the inherent technical complexity to genotype inversions in multiple samples, especially those mediated by inverted repeats or segmental duplications, they have been poorly explored so far at the population level[13].

Recently described methods using SNP genotype data can detect non- or low-recurrent genomic inversions for which recombination between the inverted and non-inverted segments has been suppressed in heterozygous carriers. The methods test for abnormal patterns of linkage-disequilibrium (inveRsion)[18], classify individuals in clusters corresponding to inversion haplotypes by pairwise identity-by-state distance matrix and multidimensional scaling (PFIDO)[15] or principal component analysis[19]. Other algorithms improve the performance of those methods by considering the existence of population admixture (invClust)[20]. We have compared the performance of haplotype and LD-based methods in previously validated inversions and found a substantial concordance in inversion calling[20]. There are at least four of such detectable common inversions by these methods, ranging in size from 0.45-4.5 Mb: inv17q21.31, inv8p23.1, inv15q24.2 and inv16q11.2. These inversions, significantly associated with effects on regional gene expression[15-17,21], are good candidates for involvement in neuropsychiatric phenotypes, since copy number variants at the same loci are causally related with neurobehavioral phenotypes[22-24].

In order to test for a possible association between ASD and the candidate inversions, the dataset of the Autism Genome Project (AGP, 2562 trios, https://www.autismspeaks.org/science/initiatives/autism-genome-project) was studied. We used available SNPs from GWAS data to genotype the four candidate polymorphic inversions using two algorithms, inveRsion and invClust, whose genotype predictions and frequencies have been previously validated in Hapmap3 and other samples[15,18,20]. While three clusters could be clearly identified by invClust at inv8p23.1 and inv17q21.31 corresponding to the homozygous standard (non-inverted, N/N), homozygous inverted (I/I), or heterozygous (N/I) subpopulations, at inv15q24.2 and

inv16p11.2 there was a clustering structure suggestive of three haplotypes with six genotypes still not well characterized at the genomic level (**Figure 1**). A high calling concordance was observed between the two methods (0.96-1), and the haplotype-alleles satisfied Hardy-Weinberg equilibrium. We found no errors in Mendelian inheritance at the inv17q21.31 and very low rate at the other three inversions, ranging from 0.1-0.4% for Europeans to 1.2-1.9% for non-Europeans (**Supplementary Table 1**). We then tested the transmission disequilibrium from parents to autistic children in AGP. No significant association or transmission distortion was observed at inv15q24.2 or 16p11.2 (**Supplementary Table 2**), but significant results were detected at inv17q21.31 and inv8p23.1 (**Table 1**).

The inverted allele (I) at inv17q21.31, also called H2, was found significantly over-transmitted to ASD probands of the AGP trio dataset (11% over the expected values of 50% per allele, corrected p=3.2e-04). Given the known phenotypic heterogeneity of ASD, we also performed a stratified analysis by the different clinical variables or subphenotypes. The most distorted (~15%) and significant over-transmission occurred in multiplex families, as well as in ASD probands who did not fulfill criteria for strict autism, being verbal and with intellectual quotient (IQ) above 80, from either simplex or multiples families (**Table 2**). We then analyzed two additional datasets with ASD trios, the Simons Simplex Collection (SSC, 2119 trios, http://sfari.org/resources/simons-simplex-collection) and the University of Miami Study on Genetics of Autism and Related Disorders (UMSGARD, 408 trios). The association of ASD with over-transmission of the H2 haplotype was not replicated in SSC (0.85%, p=0.745), and stratification analysis for subphenotypes did not yield any significant association either. In MSGARD an over-transmission of the H2 allele was also noticed, although it did not reach significance likely due to the small sample size (8.5%, p=0.216). A meta-analysis of the three datasets provided significant evidence of an overall 6.02% over-transmission of the H2 allele (p=0.0016), 5.58% (p=0.0006) considering only families of European ancestry. Although maternal transmissions were non-significantly biased in the same direction (3,9%, p=0.175), the over-transmission of H2 alleles in ASD could be considered paternal-specific (9.91%, p=5.3e-04). Despite this clear overall association of ASD with paternal H2 over-transmission, there were remarkable differences between studies with completely null results in SSC. One possible explanation for the discrepancy can be related to the different composition and

phenotype of the samples in the three studies analyzed. While AGP and also UMSGARD are enriched in multiplex families and include a significant proportion of ASD cases that do not fulfill criteria for strict autism, SSC is restricted to cases of simplex families with the great majority fulfilling criteria for strict autism. In fact, when using only AGP trios in which the case had a strict autism diagnosis, the TDT was not significant. Therefore, the H2 or I allele at inv17q21.23 appears to be associated with increased risk for a specific subphenotype of ASD with preserved verbal communication and high IQ, and is more prevalent in families with more than one case of ASD.

Over-transmission of the I-allele at the 4.5Mb inv8p23.1 was also observed consistently in the three independent ASD datasets, significant in AGP (9.59%, p=2.5e-05) and SSC (6.59%, p=0.03989) and with the same tendency in UMSGARD (11.4%, p=0.0596). A meta-analysis of the three studies provided significant evidence of an overall over-transmission by 7.1% of I-allele the most common allele in Europeans (p=2.5e-05) (**Table 1**), which is the inverted with respect to the reference genome[15]. The contribution of maternal and paternal transmissions was similar and in the same direction, with no gender bias. Stratification analysis using subphenotypes only yielded a significant increment of the association in the SSC dataset for verbal ASD with low IQ (16% over-transmission, p=0.0059) (**Table 2**).

We then performed a case-control study for ASD to further validate the findings. Since no controls other than the parents and siblings were available in any of the three studies analyzed, we used the controls of the Genetic Association Information Network (GAIN and non-GAIN) study, consisting on tested adults with no psychiatric conditions. In other to minimize the known population stratification[25] we restricted the analyses to the samples of European ancestry clustered together after principal component analysis with selected SNPs[26], a total of 4358 cases and 2392 controls. There was a significantly increased risk of ASD associated with the H2 allele of inv17q21.31 in the entire ASD dataset, with anodds ratio (OR) of 1.14 (95% confidence interval, CI=1.00-1.36) (**Table 3**). As expected, the OR was higher in AGP (OR=1.22, 95%CI=1.10-1.34) and with specific subphenotypes (**Table 3**), while there was no association in SSC (**Supplementary Table**), reinforcing the data obtained with TDT. Although inversion haplotypes have been further defined into several subtypes based on complex structural

variation at the segmental duplications and with tagging SNPs[25,27], their frequencies were too small to differentiate whether a subhaplotype was responsible for the association. For inv8p23.1, the case-control study also detected a positive association of the I-allele with ASD, with an OR of 1.22 (95%CI=1.10-1.36)on meta-analysis of the three datasets (**Table 3**). The population attributable risk of ASD was 5% for inv17q21.31 and 6.4% for inv8p23.1, with a combined attributable risk for inversion alleles of 11.08%.

We also performed a case-control study for SSD with identical filtering of the GAIN and non-GAIN datasets, using a total of 1921 SSD cases and the same 2392 controls of European ancestry. Interestingly and in contrast with ASD, the H2 allele at inv17q21.23 was found significantly associated with protection for SSD (OR=0.86, CI=0.78-0.96, p=0.0007) (**Table 3**). However, no significant association of inv8p23.1 with SSD was observed (**Table 3**).

In order to increase the number of meiotic transmissions unrelated to ASD studied, we tested these two inversions on other independent trio datasets, including studies on attention deficit hyperactivity disorder (IMAGE) and asthma (SHARP). A milder transmission distortion with over-transmission of the same I-alleles was observed in all datasets. The meta-analysis of all non-ASD trios (normal siblings in SSC, IMAGE and SHARP, a total of 2459 trios) revealed a non-significant tendency for inv17q21.31 (4.25%, P=0.11) and a significant over-transmission of the I-allele for inv8p23.1 (5.89%, p=0.009). Additional analyses searching for independent association and transmission distortion in ASD with SNPs within the two inversions yielded similar results as inversion haplotypes, consistent with the LD scores, and higher in ASD. In addition, TDT analysis using highly informative SNPs (minor allele frequency >0.4) located at 4 loci per chromosome 8 and 17 (distal and proximal p arm, proximal and distal q arm) resulted in transmission values in equilibrium between -2.9% to 3.7% of the expected 50%. Therefore, we detected no signal near centromeres or telomeres, discarding a general distortion of the allelic transmission or complete malsegregation affecting either chromosome 8 or 17. Therefore, these two inversions shower some transmission distortion to a less extent in the general population of the same alleles associated with increased risk of ASD, significant for inv8p23.1. The documented transmission distortion can be due to meiotic drive, competition among gametes, or

viability selection. The lack of signal near centromeres or telomeres, the regions most susceptible to female-specific meiotic drive, suggest any of the other two mechanisms as more likely.

The 0.9Mb inv17q21.31 and the 4.5Mb inv8p23.1 have already been extensively studied at the structural, functional and population levels, as well as in several disorders. Inv17q21.31 has two well-defined inversion haplotypes (H1/H2), as well as subhaplotypes[25,27]. The world-wide more common H1 haplotype (70-80% in EUR, >90% in Africans or Asians)[27] has been reported as a risk factor for supranuclear palsy, corticobasal degeneration, Alzheimer and Parkinson diseases[28-30]. The H2 haplotype is under positive selection in Icelanders, with carrier females having more children and higher genome-wide recombination rates than noncarriers[14]. SNPs at the *CRHR1* gene in the H2 haplotype predict a better response to inhaled corticosteroid in asthma[31]. H2 is a protective factor through gene-environmental interaction for sexual abuse-associated alcohol dependence[32] and variants in the region were associated with cranial volume in infants of the general population in GWAS[33]. The haplotypes have been shown to correlate with gene expression in several tissues, both for multi-copy genes at the segmental duplications (increase of *LRRC37A4* and decrease of *LRRC37A* in H1*)* and for single copy genes within the inversion interval (increase of *PLEKH1M* and *MAPT*, and decrease of *MGC57346* and *CRHR1* in H1, in specific brain areas)[17]. In addition to the accumulated sequence variations between inversion states, differential methylation patterns at both alleles have also been proposed to underlie the remarkable expression differences between inversion alleles at inv17q21.31[34].

On the other hand, inv8p23.1 has also been indirectly linked with the susceptibility to autoimmune disorders such as rheumatoid arthritis and lupus[15], and the inversion is a recessive risk factor for underweight in European children[20]. Inversion alleles have remarkable population stratification, following a clinal serial founder effect distribution model[15]. I-allele frequency is 70-80% in Africans, 54-60% in Europeans and <15% in Asians. The influence of inv8p23.1 on local gene expression has been reported in several data sets with transcriptomic data from lymphoblastoid cell lines and post-mortem liver samples[15]. Inv8p23.1 was robustly associated with expression levels of the following genes: *BLK*, *PPP1R3B*, *XKR6*, *FAM167A*, and *CTSB*. Since the associations were stronger with single SNPs than with the inversion, the data suggested that the

expression-related quantitative trait loci at inv8p23.1 are primarily mediated by maintained allelic configurations with functional SNPs.

In order to further characterize the functional consequences of the inversion polymorphism and define candidate genes for the associated phenotypes, we re-analyzed gene expression levels in two datasets from brain regions. We validated the differential allelic expression in brain regions for inv17q21.3. In particular, *CRHR1* expression directly correlated with the number of H2 alleles, mainly in the frontal cortex and cerebellum, while *MAPT* expression was in the opposite direction (**Figure 2** & **Supplementary Table**). A tendency in the same direction of *MAPT* was observed for *KANSL1* and *ARHGAP27*. Overexpression of *CRHR1* in cerebellum and frontal cortex was found associated with the H2-I allele of inv17q21.31. *CRHR1* encodes a G-protein coupled receptor for several neuropeptides of the corticotrophin releasing hormone family that are major regulators of the hypothalamic-pituitary-adrenal axis. This axis underlies both adaptive and maladaptive responses to stress and may be an important marker of childhood vulnerability to psychopathology. Genetic variation at *CRHR1* has been shown to interact with early life stress to predict adult depression and to regulate the effect of childhood maltreatment on cortisol responses to stimulatory tests[35]. Haplotypes at the *CRHR1* LD region have also been significantly associated with cortisol reactivity measured in saliva after standardized stress tasks in children[36]. In addition, reducing CRHR1 levels in mice with duplication of the *Mecp2* gene showed beneficial effects on the anxiety and autism-like features of these mice[37]. Therefore, overexpression of *CRHR1* in H2 carriers at several brain regions might confer susceptibility to ASD through facilitating the deleterious effect of prenatal and early infancy stressful events mediated by cortisol reactivity on neuronal synapses. This is consistent with the phenotype of 17q21.31 microduplications have recently been reported in children with mild intellectual disability and ASD features[26].

Additionally, inv8p23.1 showed a significant correlation with several genes of uneven distribution in brain regions. The I-allele was correlated with overexpression of *MSRA* in cerebellum and down-regulation in pons, increased expression of *SOX7* in cerebellum and pons, down-regulation of *FDFT1* in cerebellum and frontal cortex, down-regulation of *CTSB* in cerebellum, and overexpression of *FAM167A* in frontal cortex and pons (**Figure 3** & **Supplementary Table**). None of these genes has been previously

associated to ASD although *CTSB*, encoding a lysosomal cysteine proteinase involved in the proteolytic processing of amyloid precursor protein, has been related with anxiety, depression-like and emotionality in mice[38]. Since there is accumulating evidence of a relevant role for immune dysregulation in the pathogenesis of ASD, it is possible that an increased susceptibility to autoimmune disorders mediated by inv8p23.1 could also underlie an increased risk for ASD[39]. Interestingly, an increased risk of ASD has been documented in children born to mothers with systemic lupus erythematosus[40].

In summary, we have found two common polymorphic inversions of the human genome, unevenly distributed among human populations due to genetic drift or adaptive selection, that are susceptibility factors for ASD, with inv17q21.23 also associated with SSD. Interestingly, risk alleles for ASD are also being over-transmitted in the European general population, suggesting that both inversions are undergoing adaptive selection. Inversion alleles at both loci correlate with differential expression in specific brain areas and blood cells of candidate genes that might converge with other environmental risk factors. Our data suggest that common polymorphic inversions may explain part of the missing heritability for other multifactorial disorders.

Acknowledgement

This work was supported by the Spanish Ministry of Science and Innovation (MTM2011-26515) and Statistical Genetics Network - GENOMET (MTM2010-09526-E).

References

1- Abecasis GR et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

2- Altshuler DM et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58.

3- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11, 446–450.

4- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747–753

5- Welcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720.

6- Rossignol, D. a, & Frye, R. E. (2012). A review of research trends in physiological abnormalities in autism spectrum disorders: immune dysregulation, inflammation, oxidative stress, mitochondrial dysfunction and environmental toxicant exposures. Molecular Psychiatry, 17(4), 389–401. doi:10.1038/mp.2011.165

7- Klei, L., Sanders, S. J., Murtha, M. T., Hus, V., Lowe, J. K., Willsey, a J., … Devlin, B. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Molecular Autism, 3(1), 9.

8- Escudero I, Johnstone M. (2014) Genetics of schizophrenia. Curr Psychiatry Rep. 16 (11):502.

9- Marta Codina-Solà, Benjamín Rodríguez-Santiago, Aïda Homs, Javier Santoyo, Maria Rigau, Gemma Aznar-Laín, Miguel del Campo, Blanca Gener, Elisabeth Gabau, María Pilar Botella, Armand Gutiérrez-Arumí, Guillermo Antiñolo, Luis Alberto Pérez-Jurado, Ivon Cuscó. (2015).

Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. Molecular Autism, 6(1), 21.

10- Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, et al. (2012) Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism, 3:9

11- Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. (2014) Most genetic risk for autism resides with common variation. Nat Genet. 46:881–5.

12- Chaste P, Klei L, Sanders SJ, Hus V, Murtha MT, Lowe JK, et al. (2015) A genome-wide association study of autism using the Simons Simplex Collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? Biol Psychiatry, 77(9):775-84.

13- Feuk L (2010). Inversion variants in the human genome: role in disease and genome architecture. Genome Medicine, 2, 11.

14- Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, et al (2005). A common inversion under selection in Europeans. Nat Genet, 37(2), 129–37.

15- Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Gen Res*, *22*(6), 1144–53.

16- González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, et al. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *Am J Hum Genet*, *94*(3), 361–72.

17- De Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, Ophoff RA. (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics*, *13*, 458.

18- Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR (2012). Identification of polymorphic inversions from genotypes. *BMC Bioinformatics*, *13*(1), 28.

19- Ma J, Amos CI. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis.

20- Cáceres A, González JR. (2015) Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Research*, *43*(8), e53-e53

21- Cáceres A., Esko T, Pappa I, Gutiérrez A, Lopez-Espinosa, MJ, Llop S et al. Association of inversion-related haplotypes at 15q24.2 with children's intelligence. Submitted

22- Grisart B, Willatt L, Destrée A, Fryns JP, Rack K, de Ravel T, Rosenfeld J, Vermeesch JR, Verellen-Dumoulin C, Sandford R. (2009) 17q21.31 microduplication patients are characterised by behavioural problems and poor social interaction. J Med Genet. 46(8):524-30.

23- Filges I, Sparagana S, Sargent M, Selby K, Schlade-Bartusiak K, Lueder GT, et al (2014). Brain MRI abnormalities and spectrum of neurological and clinical findings in three patients with proximal 16p11.2 microduplication. Am J Med Genet A. 164A(8):2003-12.

24- Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. (2011) A copy number variation morbidity map of developmental delay. Nat Genet. 43(9):838-46.

25- Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012). Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genetics*, *44*(8), 881–5.

26- Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mut*, *30*(1), 69–78.

27- Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, et al. (2012) Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet. 44(8):872-80.

28- Webb A, Miller B, Bonasera S, Boxer A, Karydas A, Wilhelmsen KC. (2008) Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. Arch Neurol. 65(11):1473-8.

29- Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC,et al (2015) A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry. Mar 17. doi: 10.1038/mp.2015.23. [Epub ahead of print]

30- Zabetian CP, Hutter CM, Factor SA, Nutt JG, Higgins DS, Griffith A, et al. (2007) Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease. Ann Neurol. 62(2):137-44.

31- Nelson EC, Agrawal A, Pergadia ML, Wang JC, Whitfield JB, Saccone FS, et al. H2 haplotype at chromosome 17q21.31 protects against childhood sexual abuse-associated risk for alcohol consumption and dependence. Addict Biol. 2010 Jan;15(1):1-11.

32- Tantisira KG, Lazarus R, Litonjua AA, Klanderman B, Weiss ST. (2008) Chromosome 17: association of a large inversion polymorphism with corticosteroid response in asthma. *Pharmacogenet Genomics* 18, 733–737.

33- Taal HR, St Pourcain B, Thiering E, Das S, Mook-Kanamori DO, Warrington NM, et al. (2012) Common variants at 12q15 and 12q24 are associated with infant head circumference. Nat Genet. 44(5):532-8.

34- Li Y, Chen JA, Sears RL, Gao F, Klein ED, Karydas A, et al. (2014). An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. *PLoS Genetics*, *10*(3), e1004211.

35- Tyrka AR, Price LH, Gelernter J, Schepker C, Anderson GM, Carpenter LL. (2009) Interaction of childhood maltreatment with the corticotropin-releasing hormone receptor gene: effects on hypothalamic-pituitary-adrenal axis reactivity. Biol Psychiatry 66(7):681-5.

36- Sheikh HI, Kryski KR, Smith HJ, Hayden EP, Singh SM. Corticotropin-releasing hormone system polymorphisms are associated with children's cortisol reactivity. Neuroscience. 2013 Jan 15;229:1-11.

37- Samaco RC, Mandel-Brehm C, McGraw CM, Shaw CA, McGill BE, Zoghbi HY. Crh and Oprm1 mediate anxiety-related behavior and social approach in a mouse model of MECP2 duplication syndrome. Nat Genet. 2012; 44(2):206-11.

38- Czibere L, Baur LA, Wittmann A, Gemmeke K, Steiner A, Weber P, et al. (2011) Profiling trait anxiety: transcriptome analysis reveals cathepsin B (Ctsb) as a novel candidate gene for emotionality in mice. PLoS One 6(8):e23604.

39- Estes ML, McAllister AK. (2015) Immune mediators in the brain and peripheral tissues in autism spectrum disorder. Nat Rev Neurosci. 16(8):469-86.

40- Vinet É, Pineau CA, Clarke AE, Scott S, Fombonne É, Joseph L, Platt RW, Bernatsky S. Increased Risk of Autism Spectrum Disorders in Children Born to Women with Systemic Lupus Erythematosus: Results from the OSLER Cohort. Arthritis Rheumatol. 2015 Aug 28.

Methods

We called the inversion genotypes on a number of data sets. We were granted permission from dbGAP (Database of Genotypes and Phenotypes) (http://www.ncbi.nlm.nih.gov/gap) and the Simons Foundation Autism Research Initiative (SFARI) to download the genotypes and clinical information from multiple studies. We also used the HapMap3 genotypes downloaded from www.hapmap.org. The total number of individuals studied was 12,835 (Supp Table)

Datasets

Autism Spectrum Disorders

Autism Genome Project (AGP) (dbGap phs000267.v4.p2): International effort collecting autism families for ongoing genetic studies. Probands were diagnosed and classified using the Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) instruments. A total of 653 cases are from multiplex families (at least two individuals first to third degree relatives receiving validated ASD diagnoses), 715 cases are from simplex families (only one individual with ASD among first to third degree relatives), and for 1,236 cases the familial status was unknown. Genotype data of 2,562 trios (2,258 trios of European ancestry) had been obtained with the Human1M-Duo BeadChip SNP array (Illumina).

Simons Simplex Collection (SSC): Characterized sample of 2,119 simplex families with a proband diagnosed and classified as ASD using the ADI-R and ADOS tools. Each family was composed by trios or quartets, with a non-affected sibling available in 1,631 families. All samples had been genotyped with one of the following Illumina arrays with more than 1 Million SNPs: 1Mv1, 1Mv3 or Omni2.5.

University of Miami Study on Genetics of Autism and Related Disorders (UMSGARD) (dbGap phs000436.v1.p1): composed by probands with a clinical diagnosis of autism, Asperger syndrome or pervasive developmental disorder from the University of Miami study on autism disorders. There is available genotype data of 422 unrelated trios, 391 of them of European ancestry.

Schizophrenia Spectrum Disorders

The Genetic Association Information Network (GAIN and non-GAIN) dbGAP phs000021.v3.p2 and phs000167.v1.p1. These studies are part of the Molecular Genetics of Schizophrenia (MGS) genome wide association study (GWAS) of 3,972 cases (2,686 EA and 1,286 AA) and 3,629 controls (2,656 EA and 973 AA), comprised of European ancestry (EA) and African American (AA) samples. The genotype of about half of the EA sample and almost all of the AA was done under the auspices of the Genetic Association Information Network (GAIN), while the remainder half of is referred as the nonGAIN sample. Both were genotyped with the Affymetrix 6.0 platform at the Broad Institute.

Attention Deficit Hyperactivity Disorder (ADHD)

International Multi-Center ADHD Genetics Project (IMAGE) (dbGaP: phs000016.v2.p2). SNP genome-wide association scan of 958 parent-child trios from the International Multisite ADHD Genetics (IMAGE) project. Probands were children aged 6-17 years with any subtype ADHD, IQ above 70, free of known single-gene disorders, neurological disease and damage, and not meeting criteria for autism or Asperger's syndrome. Genotype information of 924 trios was obtained with the Perlegen-600K array.

Asthma

The SNP Health Asthma Resource Project (SHARP) (dbGAP phs000166.v2.p1) is conducting a genome-wide analysis in adults and children with asthma. Each of the childhood asthma studies has a majority of children participating as part of a parent-child trio. The genotypes of the available 435 trios had been obtained with the Affymetrix 6.0 platform.

Quality control metrics for all the trio-based studies was based on genotypic sex, Mendelian errors and familial relationships using identity by descent and genotyping quality. See dbGap accession number links for further information about study designs.

Expression datasets

To assess the correlation with inversions and gene expression, we used several available studies with genotype and expression data from blood and brain tissues. Data of the

transcriptomic analysis of blood samples of CEU individuals were obtained from the European Bioinformatics Institute at EMBL (http://www.ebi.ac.uk/arrayexpress/), generated through the project E-MTAB-198. For this analysis only parents were analyzed and individuals NA12282 and NA12283 were removed since they are grandparents of family 1421. Therefore, 105 individuals were included in CEU population.

We re-analyzed whole-genome genotyping and expression analysis on a series available studies, of 193 neuropathologically normal human brain samples using the Affymetrix GeneChip Human Mapping 500K Array Set and Illumina HumanRefseq-8 Expression BeadChip platforms (GEO: GSE15745) (Myers et al 2007).

European ancestry assessment

European ancestry of the selected samples were determined using a subset of 129,454 SNPs shared among the different datasets performing multidimensional scaling method (MDS) with 247 *hapmap3* samples for each platform, to cluster European samples. We then projected the data from the different datasets separately onto these MDS, and defined as non-European all individuals whose points fell outside the defined cluster. Some studies already assessed to be from European ancestry, and were contrasted with the results. Additionally, all platforms were merged together using 125 available SNPs in order to segregate the different populations. With this information we assessed the European ancestry of the samples.

In order to validate the accuracy of our control method for stratification we randomly selected 63 samples of HapMap3 panel with different ethnic groups (Yoruban, Japanese, Toscans and Americans with European ancestry). Blind analysis of the clustered genotypes perfectly classified the ethnic groups in all cases (**Supplementary figure**).

Inversion genotyping

Two algorithms were applied in order to identify potential genomic inversions, following the same strategy performed in previous studies. The first algorithm, inveRsion, is based on differences in LD between SNP blocks across predicted inversion breakpoints. This algorithm searches for possible breakpoints, based on

differences of Bayes Information Criterion (BIC). When BIC is greater than zero, it suggests that the chromosomes of some individuals are more likely than not to harbor an inversion in between the tested interval. The second algorithm, invClust, is based on *PFIDO* and uses multivariate analysis of SNPs within a region to predict inversion-related haplotypes. Then *invClust*, automatically applies a clustering k-means method to identify the main haplotype groups in the dataset. Theoretically, an inversion will cluster all individuals into three groups (homozygous for the inverted and non-inverted alleles and the heterozygous ones). Posterior probability (e.g. confidence) for each cluster is provided by the algorithm. Therefore, in order to minimize genotyping errors, we only considered for subsequent analyses the samples with an *invClust* confidence >0.99 (based on the number of samples with undefined clustering). To infer the error rate of inversion calling for each of the 4 inversions, we used a dataset of 2200 European trios and checked the Mendelian transmissions of predicted inversion haplotypes.

In order to define the haplotypic substructure at the inv17q21.31, we analyzed the genotypes at SNPs known to be in complete LD with the defined structures[25,26], define the H1 and H2 subhaplotypes, and run again a TDT with the AGP trios.

Genetic Association Analyses

For transmission disequilibrium testing (TDT), we analyzed only complete trios (mother, father and proband and/or sibling if available) using *snpStats* R package as well as manual calculation of TDT using the genotyped inversion alleles. The TDT is a McNemar's test based on the the binomial distribution. The null hypothesis corresponds to H_0: $p_{A1}=p_{A2}=1/2$, where p_{A1} is the probability of transmitting the A_1 allele (or N) and p_{A2} is the probability of transmitting the A_2 allele (or I). Test statistic: $X = (b - c)^2/(b + c)$, where b and c are the numbers of observed transmissions of the N and I alleles, respectively. We performed the TDT in all datasets for all parental transmissions, separately analyzing paternal transmissions and maternal transmissions. When all members of a trio were heterozygous at an inversion and the allele transmitted by each parent could not be identified, 0.5 was added to both b and c when calculating the paternal and maternal test statistics. This biases the test statistic toward the null and produces estimates of allele transmission rates that are closer to 50%. We considered loci to be over-transmitted when reached a particular significance threshold ($P<0.05$), and there would be a parental-specific over-transmission when the significance

threshold is only derived from maternal or paternal over-transmission. TDT is not sensitive to population stratification; however, heterogeneity in ancestry could dilute the signal of a geographically restricted segregation distorter or selected allele. Therefore, TDT was performed using all samples, and also selecting only Europeans.

To investigate the overall transmission of other regions of chromosome 8 or 17, we selected 8 highly informative SNPs (minor allele frequency >0.4) located at 4 loci per chromosome (distal and proximal p arm, proximal and distal q arm) and performed TDT analysis as above.

Case-control association tests were also performed using inversion genotypes, selecting only those homogeneous European samples (probands and controls) according the population stratification values.

Statistical analysis

Population stratification was assessed using principal component analysis (PCA). Linkage Disequilibrium (measured using R squared statistic) between the inversion genotype and the SNPs in the region of interest was computed using "snpStats" package.

We used the goodness-of-fit χ^2 test to determine if the paternal and maternal transmission ratio of two alleles of a certain inversion deviated from the expected ratio of 1:1,. A standard χ^2 test was used to evaluate the transmission ratio for any given inversion. Paternal and maternal allele transmission ratios were analyzed jointly and separately.

Association analysis between phenotype and inversion genotype was performed with generalized linear models (glm) from *SNPassoc* R package. All models were adjusted for the tow first principal components in order to correct for population stratification when necessary. Gender was also considered as a covariate in the models to deal with possible differences between males and females. We assumed an additive model when testing inversion genotypes effect.

For all models (TDT as well as glm) Bonferroni correction was used to adjust the *p*-values of association to consider multiple testing (four inversions) Population Attributable Risk (PAR) was computed using this formula: $PAR = P_{inv}(OR_{inv}-1) / [1 + P_{inv}(OR_{inv}-1)]$, where P_{inv} is the prevalence of the inversion (e.g., proportion who are carrying the inverted allele) and OR_{inv} is the odds ratio of disease due to the inverted allele. The combined PAR was also computed using this formula $PAR = 1 - (1- PAR_{inv17})(1- PAR_{inv8})$.

Differential gene expression analysis between inversion genotypes and genes was performed using generalized linear models implemented in the *limma* package. Surrogate variable analysis (from *sva* pacakge) was used to correct for any batch effect or hidden covariate in gene expression data.

Tables and figures

Table 1. TDT analyses for inv17q21.31 and inv8p23.1 in all trio datasets. Only samples with genotype confidence >0.99 were used. Trios: number of complete trios with genotype data. ME: Mendelian errors. IAT: inversion alleles transmitted. NAT: normal (non-inverted) alleles transmitted. %OT: percentage of over-transmission.

inv17q21.31 (C=1)	trios	ME	IAT	NAT	%OT	Pvalue
ASD all	5087	0	1442	1278	6.02	0.002
ASD EUR all	4246	0	1296	1128	5.58	0.001
AGP	2258	0	715	573	11.02	8.0e-05
SSC	1631	0	472	464	0.85	0.745
UMSGARD	377	0	115	97	8.49	0.217
SSC sibs	1363	0	400	369	4.03	0.264
IMAGE EUR	661	0	246	232	2.93	0.522
SHARP EUR	435	0	138	121	5.57	0.291
Non-ASD EUR	2459	0	784	722	4.12	0.110

inv8p23 (C>0,99)	trios	ME	IAT	NAT	%OT	Pvalue
ASD all	4808	6	1869	1620	7.13	2.5e-05
ASD EUR all	4165	1	1673	1418	8.24	1.8e-04
AGP	2179	0	875	719	9.79	9.0e-05
SSC	1626	1	647	579	5.54	0.052
UMSGARD	360	0	151	120	11.44	0.060
SSC sibs	1266	1	499	456	4.50	0.164
IMAGE EUR	632	0	348	306	6.42	0.100
SHARP EUR	389	0	168	140	7.37	0.111
Non-ASD EUR	2287	0	1015	902	5.89	0.010

Table 2: TDT categorizing AGP samples according to different subphenotypes. SPC: phenotype within the spectrum. STR: strict definition of autism

inv17q21.31 (C=1)	variable	category	trios	IAT	NAT	%inc	OR	LoCI	UpCI	P-value
AGP	ethnicity	EUR	2258	715	573	11.02	1.24	1.09	1.42	8e-5
	family	multiplex	653	189	139	15.24	1.35	1.04	1.77	0.006
	verbal	yes	1761	536	433	10.62	1.23	1.06	1.44	0.001
	IQ	high>80	1033	331	243	15.33	1.36	1.11	1.66	2.4e-4
	gender	male	2246	666	547	9.81	1.21	1.06	1.39	6.3e-4
	spectrum	yes	2437	736	598	10.34	1.23	1.07	1.40	1.6e-4
	strict	no	1106	353	268	13.68	1.31	1.08	1.59	6.5e-4

inv8p23 (C=0.99)	variable	category	trios	IAT	NAT	%inc	OR	LoCI	UpCI	P-value
AGP	ethnicity	EUR	2179	875	719	9.78	1.21	1.07	1.37	9e-5
	family	multiplex	651	252	209	9.32	1.20	0.96	1.50	0.045
	family	simplex	709	302	250	9.42	1.20	0.98	1.48	0.026
	family	unknown	1233	480	405	8.47	1.18	1.00	1.39	0.011
	verbal	yes	1755	714	556	12.44	1.28	1.12	1.47	9.3e-6
	IQ	high>80	1052	426	346	10.36	1.23	1.03	1.46	0.004
	spectrum	yes	2424	956	813	8.08	1.17	1.04	1.31	0.001
	strict	no	1101	443	358	10.61	1.23	1.04	1.46	0.003
SSC	ethnicity	EUR	1456	585	516	6.26	1.13	0.98	1.31	0.037
	verbal	yes	1309	525	460	6.59	1.14	0.97	1.33	0.038
	IQ	low<70	381	158	118	14.49	1.33	1.00	1.79	0.016
	verbal and IQ<70		263	113	71	22.82	1.59	1.10	2.28	0.001

Table 3. Association study of inv17q21.31 and inv8p23.1 with ASD and SSD, additive model

H2 at inv17q21.31	cases	controls	OR	LoCI	UpCI	P-value
AGP (EUR)	2258	2392	1.2197	1.1077	1.3431	5e-5
SSC (EUR)	1709	2392	1.0555	0.9298	1.1981	0.404
UMSGARD (EUR)	391	2392	1.1780	0.9489	1.4624	0.138
ASD all (EUR)	4358	2392	1.1382	1.0455	1.2392	0.003
SSD all (EUR)	1921	2392	0.865	0.778	0.961	0.007

I at inv8p23.1	cases	controls	OR	LoCI	UpCI	P-value
AGP (EUR)	2179	2303	1.1210	1.0406	1.2075	0.003
SSC (EUR)	1631	2303	1.1243	1.0243	1.2341	0.014
UMSGARD (EUR)	374	2303	1.1140	0.9470	1.1140	0.192
ASD all (EUR)	4184	2303	1.1210	1.0406	1.2759	0.003
SSD all (EUR)	1855	2303	0.9842	0.8998	1.0766	0.729

Figure 1. Representation of the genomic regions harboring common ancestral inversions and the individual classification by invClust. Relevant genes are represented by arrows in the transcriptional direction, and the segmental duplications are represented according to the UCSC browser. On the right, the plots show the MDS clustering by invClust of the individuals for each of the four inversions, using genotypes of the SNPs located between the segmental duplications.

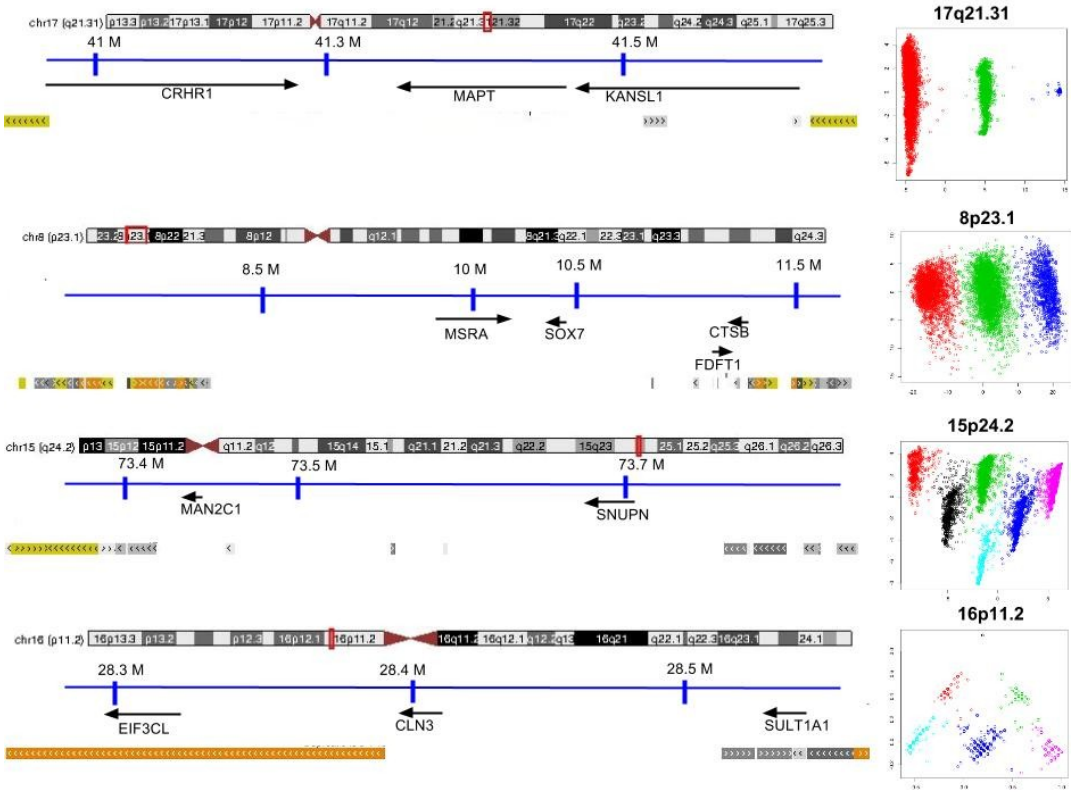
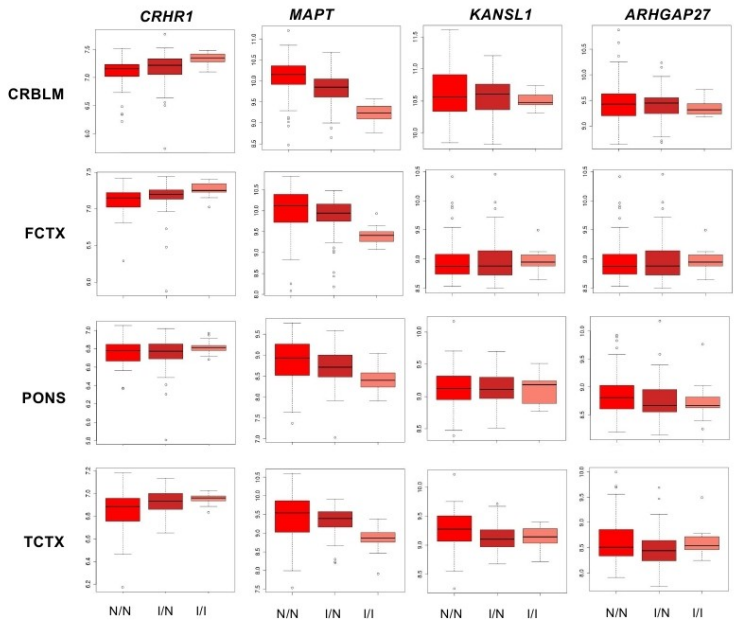
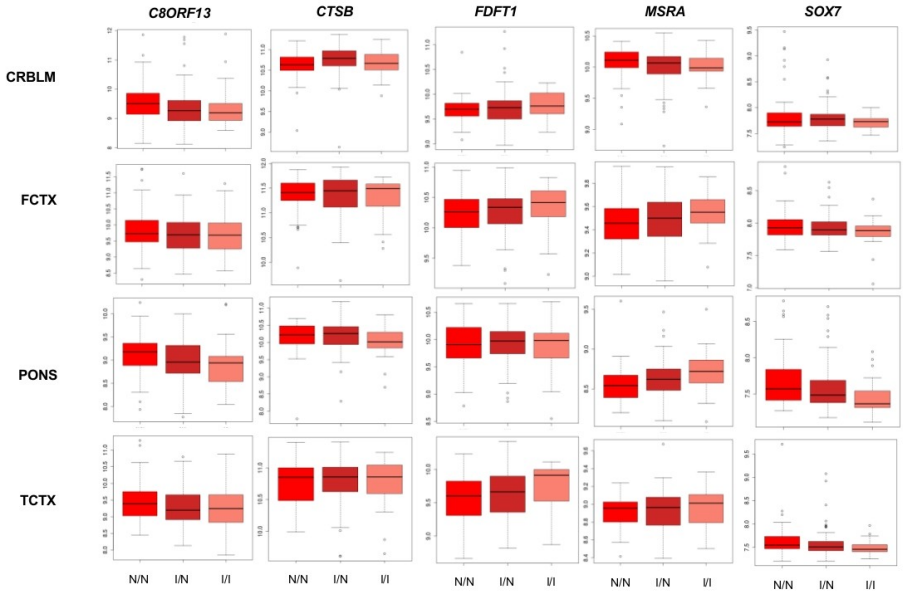


Figure 2. Correlation of inv17q21.31 genotypes (A) and inv8p23.1 genotypes (B) with local gene expression in several brain regions.


CRHR1
MAPT
KANSL1
ARHGAP27
CRBLM
FCTX
PONS
TCTX
N/N
I/N
I/I

.


C8ORF13
CTSB
FDFT1
MSRA
SOX7
CRBLM
FCTX
PONS
TCTX
N/N
I/N
I/I

5.2 Unearthed ancient haplotypes at microdeletion region 15q24.2 are linked to a novel inversion signal, brain expression of MAN2C1 and children's intelligence

Alejandro Cáceres, Tõnu Esko, Irene Pappa, Armand Gutiérrez, Maria-Jose Lopez-Espinosa, Sabrina Llop, Mariona Bustamante, Henning Tiemeier, Andres Metspalu, Peter K. Joshi, James F. Wilson, Zdenka Pausova, Tomáš Paus, Jordi Sunyer, Luis A. Pérez-Jurado, Juan R. González **Unearthed ancient haplotypes at microdeletion region 15q24.2 are linked to a novel inversion signal, brain expression of MAN2C1 and children's intelligence.**
(Submitted)

Abstract

The chromosome bands 15q24.1-15q24.3 contain a complex region with numerous structural variations that include microduplications and microdeletions, both of which have been linked to intellectual disability, speech delay and autistic features. The region is also known to harbour common inversion polymorphisms whose functional and phenotypic manifestations are unknown.

Using single nucleotide polymorphism (SNP) data, we detected a novel ~0.6Mb inversion signal in 15q24.2. We characterized three putative inversion-induced haplotypes (NI, Ia and Ib) that showed Mendelian inheritance in trio analysis and high frequency in Europeans (NI=20%, Ia=50%, Ib=30%). Worldwide population analysis revealed an African origin of the three haplotypes with significant population stratification. The three alleles strongly correlated with expression levels of *MAN2C1* and *SNUPN* in blood and brain. Homozygosity for NI correlated with over-expression of *MAN2C1* over many brain areas but the occipital cortex where Ib homozygosity was highly under-expressed. While no association of the alleles was observed for autism, we found association with verbal and non-verbal intelligence quotient (IQ) in 2,735 children of European ancestry from three independent population cohorts. Homozygosity for NI was associated with lower verbal IQ (2.5-point loss, p-value=0.008), while homozygosity for Ib was associated with 3.2-point loss in non-verbal IQ (p-value=0.0009). Common polymorphic inversion-related haplotypes at 15q24.2 may influence human intelligence most likely by regulating gene expression of local genes *MAN2C1* and *SNUPN*.

Author Summary

Rare deletions and duplications of the human genomic regions between chromosome bands 15q24.1-15q24.3 have been linked to intellectual disability and autism with verbal difficulties. Therefore, as altered genes are fundamental in the development of human cognition, their variability may also play a role in the cognitive differences of healthy children. Previous research has found inversion polymorphisms in 15q24, which are frequent in the general population. While inversions can sustain extended haplotypes that maintain genetic variability in the population, they are difficult to characterize in the lab for a large number of people. We use current bioinformatics methods to detect a novel inversion signal in the region which allow us to determine the extended haplotypes it supports. We study the African origin of the haplotype structure, its effect on the expression of local genes in the brain and the possible links to the cognitive ability of normal children. We find that different haplotypes change the expression of *MAN2C1* in different parts of the brain consistently with the differences in cognitive abilities of the children studied, and with the phenotypic characteristics of the reported microdeletions and microduplications cases. These inversion-supported haplotypes may influence human intelligence likely through the expression of *MAN2C1*.

Introduction

Inversion polymorphisms are poorly explored sources of genetic variability in the general population. They can alter gene expression by breaking genes, disrupting regulatory elements, or maintaining haplotype structures though suppression of recombination when heterozygous. A database with all reported inversions in the human genome has been recently updated [1] and includes a number of validated inversions with high frequency in the population. It has also been recently shown that the inversion status of some of these inversions can be consistently inferred from SNP data using the congruence of two bioinformatics methods [2].

The chromosome bands 15q24.1-15q24.3 harbour a complex genomic region with multiple large blocks of segmental duplications (A through E) that mediate recurrent rearrangements, including inversions, deletions and duplications of variable size and extent [3,4,5]. Both, microdeletions and microduplications of this region cause unusual facial morphology along with intellectual disability, speech delay and autistic features

[6,7]. Most reported deletions include the 1.1 Mb critical region located between blocks B and C and also the 0.6 Mb C–D region where smaller deletions have been reported in at least two patients with borderline intellectual disability [6]. Thus, while the severe core cognitive deficits of the 15q24 microdeletion syndrome are thought to be due to deletion of genes between B and C, some of the genes located between blocks C and D must also be important for normal development and behavior. Here, we have found an inversion signal in the 15q24.2 region between blocks C and D, detectable by differences in linkage disequilibrium (LD) between SNP blocks and haplotype divergence between inversion states [2,8]. Given the reported implication of gene dosage effects of this 15q24.2 region in autism and cognitive deficits, we aimed to investigate the evolutionary history, the effect on gene expression and putative influence of this inversion polymorphism on autism susceptibility and on the general cognitive ability (GCA) of children and adolescents recruited from the general population.

Results

Characterization of inversion-induced haplotypes at 15q24.2

We used two recent bioinformatics methods, inveRsion and invClust, to detect inversion signals within 15q24.1-15q24.3, see more details in the Methods section. Scanning with inveRsion SNP data of the CEU individuals the region between 70-77 Mb of chromosome 15 (hg18), we detected three positive signals (BIC>0) of LD differences (Figure 1D). One of those signals corresponded to the 0.4 Mb interval between 73.29-73.72 Mb, mapped within the 0.6Mb segment that is flanked by the C and D segmental duplications (73329051-73839599 Mb). The other two signals were obtained between segmental duplications D-E, which did not merge with higher window sizes. We did not detect a signal consistent with the previously reported 1.1 Mb inversion between blocks B and C [3,4].

We then used the multivariate method, invClust, to determine the haplotype structure of the candidate-inverted regions, defined as regions between two segmental duplication blocks. As previously reported [9], we kept the first component of a multidimensional scaling (MDS) analysis, which revealed a clear five-cluster pattern for the segment C-D (15q24.2) (Figure 1C). Other segments (A-B, B-C, and D-E) did not show a clear

haplotype substructure, and were not analyzed further as their predictions did not converge with inveRsion.

Further analysis on the CEU and YRI populations of the HapMap III (n=360) revealed that samples could be classified in six possible clusters, using the first two components of the MDS analysis (S1 Fig, S1 Table). Each cluster could represent the six possible haplotype-genotypes of three different groups, namely NI (non-inverted), Ia and Ib (inverted). The three extreme clusters would represent the homozygous individuals (Ia/Ia: cluster 1, Ib/Ib: cluster 3, NI/NI: cluster 6) and the clusters in between two homozygous groups would contain the corresponding heterozygous individuals (Ia/Ib: cluster 2, Ia/NI: cluster 4, Ib/NI: cluster 5). We found, without exception, that the three haplotypes satisfied Mendelian patters of inheritance and Hardy-Weinberg Equilibrium. We also used the SNPs of the AGP that includes 2,259 European and 303 non-European trios. We observed a very low rate of errors in Mendelian inheritance for the three predicted haplotypes: 0.4% for Europeans and 1.9% for non-Europeans (S2 Table). Thus, this result indicated that the congruence of methods was highly accurate to infer the three haplotype-genotypes in large population datasets.

Population frequency of inversion-induced haplotypes at 15q24.2

While the reference NI haplotype was the most common in YRI, the most frequent allele in the CEU population of HapMap was Ia. Frequencies were as follows: NI: 59%, Ia: 24%, Ib: 17% for YRI and NI: 24%, Ia: 49%, Ib: 27% for CEU.

We analyzed the 1000 Genomes populations to determine the global frequency of inversion-related haplotypes. We used invClust on the first two MDS components of the SNPs in the C-D region for the entire sample, and observed that the haplotypes Ia and Ib are classified in a single group with respect to NI (S2 Fig).

We computed LD r^2 between the SNPs in the 15q24.2 region and the haplotype status (NI, Ia and Ib) of each CEU in the 1000 Genome data. We found over 250 SNPs that could be used to tag a given haplotype with $r^2>0.9$ (S3 Table). We selected 26 SNPs with $r^2>0.95$ and mapped the predicted ancestral homozygous state into the MDS analysis of the CEU samples (Figure 2A, S4 Table). We observed that this hypothetical ancestral homozygous individual was heterozygous NI/Ib (group 5), indicating that the ancestral configuration was equally related to both the NI and Ib haplotypes. This

suggests that a split of the ancestral allele into NI and Ib, perhaps due to a second inversion event, could have followed after the initial generation of the inversion Ia.

Haplotype frequencies across worldwide populations showed an uneven distribution consistent with an ancestral African origin of all three haplotypes and Out of Africa Expansion (Figures 2B-2C). While the NI allele is the most frequent in Africans (64% in Kenyan) and less frequent in Latin Americans (18% in Mexican), the Ia allele is less frequent in Africans (23% in Yoruba) and common in Latin American populations (67% in Mexican). The Ia allele presented a high differentiation between distant populations (F_{ST}=0.21 between Mexican and Yoruba, Figure S3), whereas the NI allele had a lower than expected frequency in Europeans, relatively to their geographical distance from Africa (F_{ST}=0.2 between British and Kenyans). The discrepancy with demographic patterns could be explained by either selection or strong genetic drift from NI to Ib (Figures 2B and 2D). As shown in figure 2D, the distribution of the Ib allele is globally uniform, except for the higher frequency in Europeans. We found a very high F_{ST} (0.88) between Iberians and Japanese, suggesting that this allele could be under selective pressures in Europeans.

We compared the entire ~7 Mb human genomic region at 15q24 (70-77 Mb, hg18) with the orthologous region in the rat genome (genome assembly rn5.0) to establish the blocks of the synteny, using ArkMAP [10]. We observed a complex evolutionary history of the region with several inversion and transposition events with some breaks of synteny at the regions harbouring the blocks segmental duplications in the human genome (S4 Fig). We found an evolutionary inversion in *Homo sapiens* for the region C-D with respect to D-E and an evolutionary deletion between the regions B-C and C-D. In addition, an additional block flanked by segmental duplications and distal to D-E has been translocated. Therefore, the human C-D region seems to have evolved independently from its flanking intervals B-C and D-E.

Functional correlation of inversion alleles with gene expression

Previous studies have shown that the inverted regions 16p11.2, 8p23.1 and 17q21.31 can influence strongly the expression of local genes [9,10,11]. Therefore, we performed association tests between the normalized gene expression within the C-D region in 15q24.2 and the inversion haplotypes. We first analyzed the genome-wide expression data of the 882 Estonians from the EGCUT study for each specific allele (S5 Fig), and

found a significant association of local genes (S6 Fig). The expression of *MAN2C1* increased per NI allele (p-value<10^{-46}), and decreased with both Ia (p<10^{-8}) and Ib (p<10^{-7}) alleles (S5 Table). An additional significant association was found for *SNUPN*, which followed the same pattern of *MAN2C1* (NI: p-value <10^{-15}, Ia: p-value =0.0005, Ib, p-value =0.003). Using transcriptomic data in lymphoblastoid cell lines of the 105 CEU individuals of HapMap, we validated the expression pattern of *MAN2C1* with respect to the NI (p-value<10^{-4}) and Ia (p-value<10^{-5}) haplotypes (S6 Table). We then used the brain expression data of 193 control individuals (S7 Fig). In agreement with the previous analyses, we found that the NI allele was associated with higher *MAN2C1* expression in cerebral cortex (p-value=0.02), see Figure 3. However, we did not find significant associations with Ia and Ib. We also tested associations between the haplotype-genotypes and the expression of *SNUPN* in brain and validated a significant reduction per Ib allele (p-value=0.05), see Figure 4.

We analyzed expression data from the BRAINeQTL study and BRAINEAC project to investigate the regional difference of gene expression in brain, see Figure 5. For the BRAINeQTL study, we correlated the expression of *MAN2C1* in four different brain areas for 148 subjects. We found that the NI allele is associated with increments of *MAN2C1* transcription in pons (p-value=0.0004), cerebellum (p-value=0.01), frontal cortex (p-value=0.01) and temporal cortex (p-value=0.02) while homozygous for Ib had significant decrements of gene expression only in pons (p-value=0.02) and no significant association was found for the Ia allele. In the BRAINEAC data-set of 134 individuals, we selected 36 intragenic probes within *MAN2C1* and tested the correlation between the recessive models for each allele and the expression of the gene across 10 different brain regions. We found only two probes that survived Bonferroni correction within each region for *MAN2C1*. The first one was in the putamen for NI homozygocity (p-value=0.0004). The second probe confirmed the high correlation in the occipital cortex only with Ib (p-value<10^{-4}) and not for NI (p-value=0.8). We did not find significant results in the other areas of the brain.

Association of inversion-related haplotypes at 15q24.2 with GCA

As copy number changes in the region have been implicated in autism and intellectual disability, we first tested the transmission disequilibrium from parents to autistic children for the three alleles, using the Autism Genome Project (AGP) dataset of 2,259

European and 303 non-European trios. Testing the odds ratio for finding an allele "*a*" in autistic children of heterozygous parents "*a/b*", we did not find any significant transmission disequilibrium in the entire sample (NI/Ib: OR=1.04, p-value=0.5, NI/Ia: OR=0.95, p-value=0.48, Ib/Ia: OR=1.03, p-value=0.5). Similarly, analysing independently strict autism and ASD cases, we did not find any significant association.

We then inferred the NI, Ia and Ib haplotypes of the C-D region for 909 and 1,236 children from the INMA and GenR population cohorts, respectively. As in the previous analysis, we identified six clusters, corresponding to the genotypes of three possible alleles in the region (S8-S9 Fig). Clusters were numbered according with the tag SNPs in S3 Table. For the SYS cohort of adolescents, we had more individuals genotyped (children and parents) but lower density of SNPs. Because only 8 SNPs from S3 Table were available for the analysis, we could only use the first MDS component where we found a clear 5-cluster pattern (S10 Fig). The Ib homozygous were inferred as those individuals who are simultaneously non-variant homozygous for NI and Ia.

We tested the association between the homozygocity for each haplotype in 15q24.2 and IQ measures, using three recessive models for NI, Ia and Ib, to assess the contribution of each allele to the verbal and non-verbal IQ in each independent cohort. We adjusted for sex, age at test administration and first two genome-wide PCA components. Afterwards, we preformed a meta-analysis where the weights were the reciprocal of the estimated variance (Figure 6). We found the Ib allele correlated with a 3.2-point loss in non-verbal intelligence (p-value=0.0009). In addition, homozygosity for the NI haplotype was the only genetic model that correlated with verbal IQ (mean decrement of 2.25 points, p-value=0.008). As expected, for all three haplotype tests, we found high correlation between verbal and non-verbal IQ estimates (cor=0.81 for NI, cor=0.70 for Ia, and cor=0.98 for Ib). For INMA and SYS studies, we analyzed general IQ score, available for these two datasets, and found that NI is associated with a 3.6-point loss in GCA (combined p-value= 0.01), see S11 Fig.

We analyzed the 2,215 adults of ORCADES to investigate if the association in verbal IQ is also present in adults. For this specific cohort, we did not find a significant association (NI: p-value=0.2, Ia: p-value=0.5, Ib: p-value: 0.06). However, since the association with NI homozigocity was also negative and comparable to that observed in children, we found a small increment on the statistical significance in the overall meta-analysis (2.09-point loss, p-value=0.005), see S12 Fig.

Discussion

We revealed the existence of a three-haplotype structure at the 15q24.2 region between the C and D blocks of segmental duplications. The signals detected with SNP data indicate that the haplotype structures are likely sustained by a yet unobserved inversion polymorphism between blocks C and D that prevented reciprocal chromosomal exchanges in heterozygous carriers. We characterized the three haplotype groups with the convergence of two independent bioinformatics algorithms, a methodology that has been previously validated in the characterization of inversion 16p11, for which a three haplotype structure has also been found [12]. The convergence approach has also been successful in the description of three other validated inversions [2]. The very low rate of Mendelian errors in trio analyses further reinforced that the haplotype-genotype inferences are robust in large population samples. We also observed that the frequencies of inversion-related haplotypes indicate their African origin with possible selection or high genetic drift. Traces of selection on inversion-haplotypes are well established and have been observed in the few inversion polymorphisms that have been characterized so far in world wide populations [9,12,13].

In addition to the evolutionary inversion in D-E, a polymorphic inversion in B-C has been reported with low frequency in the population while another one in A-B has been predicted with paired-end-mapping [1,3,4]. Thus, the 15q24 genomic region is highly dynamic. This variable structure between individual chromosomes may add additional instability through meiotic mispairing and increased susceptibility to the reported recurrent germline rearrangements at 15q24 [6], as it has been shown in other genomic regions [14]. Given such complexity, more than one inversion configuration in size and extent is possible. That is, the orientation of the region C-D may be polymorphic and its breakpoints extend to other segmental duplication blocks for different individuals. This, in addition to the relatively small size of the single copy interval (<0.35 Mb), poses great difficulties to the experimental characterization of the polymorphic inversion by Next Generation data or cytogenetic methods such as FISH. While this could be attempted using also sequence assembly technologies (PacBio or SMRT), large scale genotyping of inversions in several individuals is currently not possible. Here, however, we give substantial evidence for complex polymorphic orientations between 15q24.1-

15q24.3, and show that the region can be analyzed with SNP data from signals that are congruent with those found in experimentally validated regions [2].

Out of the 10 single copy genes located within the C-D interval, the expressions of *MAN2C1* and *SNUPN* were consistently up-regulated at the NI allele and down-regulated at the Ib configuration, both in blood and brain tissues. *MAN2C1* has been shown to have a dual function. *MAN2C1* encodes the alpha-mannosidase, class 2C, member 1 that has been shown to regulate protein N-glycosylation and apoptosis. The N-Glycoproteome maps mainly to blood but the highest amount of organ specific N-glycosylation sites in mice has been observed in the brain [15]. As the *MAN2C1* gene is highly expressed in hippocampal formation, its differential regulation by the different haplotypes might be related to the observed differences in cognitive function. Over-expression of *MAN2C1* leads to protein underglycolysation and up-regulation of the degradation of unfolded glycoproteins [16]. The attachment of glycans to some proteins is important for their correct folding and/or stability. N-glycans cover diverse biological functions in the nervous system, ranging from the essential to the modulation of development and neural transmission, which in turn can affect plasticity and memory formation [17]. Multiple glycosylation disorders with associated neurological symptoms and impaired cognitive ability have been reported [18]. An additional function of *MAN2C1*, independent of N-glycosylation, is apoptosis signalling and tumour growth [19,20]. Down-regulation of *MAN2C1* is linked to increments in apoptosis. The apoptotic action of the gene in the nervous system remains to be directly observed. Nevertheless, an indication of such function in the brain is given by previous findings showing that *MAN2C1* is over-expressed in patients with posttraumatic stress disorder [21,22], and that such psychopathology presents reduced apoptosis associated with defects in signal plasticity [23]. Therefore, the literature indicates that both over-expression and under-expression of *MAN2C1* can have a negative impact in brain function through different signalling pathways.

SNUPN is the other widely expressed gene whose expression in blood and brain is modulated by the haplotype-genotypes. *SNUPN* encodes snurportin 1, a protein required, trough interaction with the spinal muscular atrophy protein *SMN,* for the nuclear import of snRNPs and splicing regulation of multiple genes [24]. As such, *SNUPN* de-regulation can affect the central nervous system but there is not yet evidence for a more direct relation to cognition.

Given that the 15q24.2 haplotypes comprise a region in which copy number changes are causative of autism and cognitive impairment [6], we hypothesized that these inversion-generated groups of haplotypes could also be susceptibility factors for autism and/or important genetic determinants of human cognition. We did not find any association with autism susceptibility. Among normal developing children and adolescents, however, we observed that individuals with the NI allele or homozygous for Ib/Ib are at risk of lower verbal and non-verbal IQ. The association in the meta-analysis for verbal IQ is improved with the inclusion of a large adult cohort, suggesting that the genetic effect may still be relevant later in life. Note that the nominal p-values of our findings are comparable to those obtained for gene set analysis of the largest GWAS on children's intelligence [25]. In such context, our results are highly meaningful, since our approach is akin a "candidate gene" study and require no correction for multiple comparisons. Our associations are therefore highly significant and unlikely due to chance, suggesting that the haplotype-genotypes at 15q24.2 contribute to subtle variations in general intelligence, both verbal and non-verbal, most likely through changes in the regulation of specific genes in the region.

Microduplications involving *MAN2C1, SNUPN* and/or other genes within C-D have been associated with autistic features and language problems while their haploinsufficiency has been related to intellectual disability. Such gene dosage associations fit exactly with our observations in normal developing population showing that 1) over-expression of *MAN2C1* is related to NI, which in turn is linked with lower verbal IQ and 2) under-expression of *MAN2C1* is related to Ib which is linked to non-verbal IQ. Remarkably, up-regulation of *MAN2C1* at NI is stronger in the frontal and temporal cortex, where language and high cognitive functions are processed, while down-regulation of Ib is most prominent in the occipital cortex, where it could affect processing of visual stimuli and thus influence non-verbal IQ. A reading of our data consistent with *MAN2C1*'s literature further suggests that NI could be linked to underglycolysation while Ib to mitochondrial-mediated apoptosis. More generally, our analyses indicate that the "where and how" a genes express in the brain is important for the interpretation of the associations between structural variations and cognitive phenotypes.

Materials and Methods

Inversion Detection and Calling

We used dense SNP data to detect polymorphic inversions in the chromosome band 15q24.1-15q24.3, prone to microdeletions and microduplications. While the imprints of inversions on nucleotide variation can be complex, it has been shown that the convergence of two different algorithms on a single prediction can adequately characterize well known inversions [2,12]. The first algorithm, inveRsion, is based on differences in LD between SNP blocks across inversion breakpoints [8]. This method allows one to search for inversion signals without previous knowledge of the breakpoints. A positive signal of an inversion is given by the difference of Bayes Information Criterion (BIC) which, if greater than zero, indicates that the chromosomes of some individuals between the tested interval are more likely to be inverted than not. We first ran the inveRsion algorithm using 0.4 Mb window sizes on the 70-77Mb (hg18) interval of the 15q24.1-15q24.3 genomic region, using the SNP genotypes of 180 CEU individuals from HapMap III (http://hapmap.ncbi.nlm.nih.gov/). The second inversion detection algorithm is based on the multivariate analysis of the SNPs within the inverted region, which detects the underlying haplotype structure induced by the inversion event. In particular, we applied a multidimensional scaling (MDS) analysis of the SNP genotypes, keeping the first two eigen-components [9]. This analysis requires the knowledge of the breakpoints of the region, therefore we used as candidate inversion brakepoints the internal limits of the segmental duplication blocks in the genomic interval that mediate the different recurrent rearrangements found in the region. To detect the inverted-related haplotypes, we used a mixture model classification, invClust [2]. We then applied a clustering k-means method to identify the groups in the data produced by multiple allelic haplotypes. We also used data from the 1000 Genomes (http://www.1000genomes.org/) including 1,091 individuals from 14 populations to determine the global distribution of the inverted allele. We also inferred the ancestral haplotype state by including a hypothetical individual who was homozygous for all the ancestral SNPs.

Expression Datasets

We analyzed gene expression levels in RNA from peripheral blood obtained in Estonian Gene Expression Cohort (EGCUT) (http://www.biobank.ee/). This cohort is composed

of 1,074 randomly selected Estonian individuals (37+/-16.6 years; 50% females) from 53,000 subjects in the Estonian Genome Center Biobank, University of Tartu. Whole-Genome gene-expression levels were obtained by Illumina HT12v3 arrays according manufactures protocols. DNA was genotyped with Human370CNV array (Illumina). The final sample size with both genotype data and gene expression data was 882 individuals. We tested genome-wide association between gene expression and inversion status on this large cohort.

Data of the transcriptomic analysis of RNA from lymphoblastoid cell lines of CEU individuals were obtained from the European Bioinformatics Institute at EMBL (project E-MTAB-198) (http://www.ebi.ac.uk/arrayexpress/), to validate the findings on the previous dataset. For this analysis only parents were selected, removing grandparents NA12282 and NA12283 of family 1421. The final sample analysed here included 105 individuals.

Gene expression data, Myers et al. 2007 [26], was then analyzed in the specific C-D region for brain cortex. The data comprises 194 samples (with one removed after quality control) from the cerebral cortex of neurpathologically normal brains and was obtained with the Illunima HumanRefseq-8 Expression BeadChip. For SNP genotyping, the Affymetrix GeneChip Human Mapping 500K Array Set was used. The expression and genotype data was downloaded from the website for The Laboratory of Functional Neurogenomics (http://labs.med.miami.edu/myers/). *MAN2C1* transcript probe was GI_6631092.

For specific regions in the brain cortex, two gene expression data-sets were obtained from Gibbs et al 2010 [27] and BRAINEAC (http://www.braineac.org/), and analyzed for *MAN2C1*, the gene that was found significant in all previous analyses. The first data-set (BRAINeQTL, dbGap accession number: phs000249.v1.p1) contains gene expression for 148 subjects in pons, cerebellum, temporal cortex and frontal cortex. The second data-set consists on expression and genotype data for 134 subjects in 10 different brain regions: White matter, cerebellum, medulla, hippocampus, putamen, substatia nigra, thalamus, occipital cortex, temporal cortex and frontal cortex. We downloaded data for 36 intragenic probes within *MAN2C1*.

Autism Dataset

The Autism Genome Project (AGP) consortium represents an international effort collecting autism families for ongoing genetic studies. Cases were classified using the Autism Diagnostic Interview-Revised and Autism Diagnostic Observation Schedule instruments. We were granted permission to access genotype data of 2,563 trios obtained with the Human1M-Duo BeadChip (llumina) (ref: phs000267.v4.p2) (http://www.ncbi.nlm.nih.gov/gap). This dataset was used to define the Mendelian inheritance of inversion-haplotypes at 15q24.2 and to test their transmission disequilibrium in autism or autism spectrum disorder (ASD).

General Intelligence Datasets

For association tests between IQ and inversion-related haplotypes, we used data from three independent cohorts of children and adolescents recruited from three different general populations: the INfancia y Medio Ambiente (INMA) Project, the Generation R (GenR) and the Saguenay Youth Study (SYS). All participants (or their parents in case of minors) gave informed consent and local ethics committees approved each cohort study.

INMA: INMA is a network of population-based birth cohorts across Spain [28] (http://www.proyectoinma.org/) established to study the impact of environmental pollutants on child development. We used a subsample of 909 genotyped children with a European ethnic origin, selected from 2042 individuals from the Menorca, Sabadell and Valencia cohorts. Children's mothers were screened and recruited at the first-trimester ultrasound scan between 1997 and 2006. Genotyping was performed with HumanOmni1-Quad-BeadChip (Illumina). The McCarthy score of children's abilities (MSCA) was administered at 4-5 years of age in 1302 children, 903 of which were genotyped and had good quality testing, excluding incomplete administration, illness or tiredness. We selected the verbal and perceptive-performance scales of the MSCA.

GenR: This is a population-based cohort that at present covers fetal life to childhood of nearly 10,000 individuals born between 2002 and 2006 in Rotterdam, The Netherlands [29]. The data we used from this study comprised of genotypes from 1,236 European descendent children, obtained with Human610-Quad-BeadChip (Illumina). Intelligence measures were obtained at 6 years of age (±0.3 years). Verbal IQ was assessed on 821 individuals using a subset of the Dutch battery [TaaltestvoorKinderen (TvK)], and nonverbal IQ was tested on 871 children using two subsets of the Dutch nonverbal

intelligence test [Snijders-Oomen Niet-verbale intelligentie Test-Revisie (SON-R 2 ½ -7)] suited for children of 2.5-7 years of age. These two subsets tap into visuo-spatial abilities and abstract reasoning and their sum highly correlates with the full SON-R IQ battery.

SYS: We used data from the Saguenay Youth Study, established to study brain and cardio-metabolic health in 1,024 adolescents, 12 to 18 years of age (15±3.5), recruited and assessed in the Saguenay Lac-Saint-Jean region, Canada, between 2003 and 2012 [30]. Genotypes of 1,953 individuals, including parents, were obtained with Human610-Quad and HumanOmniExpress-BeadChips (Illumina), only SNPs present in both chips were analyzed. We performed association tests on a subset of 1,011 adolescents who were tested with Wechsler Intelligence Scale for Children III (WISCIII). We used the verbal and performance components of the full test.

ORCADES: We also studied whether the genetic associations persisted in adults, using a cohort of 2,215 individuals with mean age of 54 years (± 15). The Orkney Complex Disease Study is a family-based study in the isolated Scottish archipelago of Orkney. Genetic diversity in this population is decreased compared to Mainland Scotland, consistent with the high levels of endogamy historically. Fasting blood samples were collected and over 500 health-related phenotypes and environmental exposures were measured in each individual. Cognitive verbal fluency was measured with the Mill Hill vocabulary scale.

While different IQ measures were obtained at different age groups (4, 6, between 12 and 18, and 54 years) in each of the cohorts, it has been shown that different test batteries for GCA are highly correlated [31], and that the genetic influence on intelligence increases with age [32]. All IQ measures were standardized to mean 100 and standard deviation of 15. As general IQ measures were not available in all cases, we analyzed verbal and non-verbal IQ separately.

Author contributions

AC performed the inversion detection, evolutionary analysis, association tests and helped draft the manuscript. TE and AM oversaw the ECGUT data collection. IP performed genetic associations in the GenR cohort. AG performed analysis on AGP data. MJLE and SL oversaw INMA-Valencia data collection. MB oversaw the

genotyping of INMA children. HT oversaw GenR data collection. ZP and TP oversaw SYS data collection. PKJ and JFW oversaw ORCADES data collection. JS oversaw INMA data collection. LAPJ and JRG oversaw the study, helped run analysis and draft the manuscript. All authors critically revised the manuscript and approved its final version.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (MTM2011-26515) and Statistical Genetics Network - GENOMET (MTM2010-09526-E), the Instituto de Salud Carlos III (FIS PI1002512 and PI1302481) and Generalitat de Catalunya (2014SGR1468). INMA was funded by grants from Instituto de Salud Carlos III (CB06/02/0041, FIS PI041436, PI081151, PI041705, and PS09/00432, FIS-FEDER 03/1615, 04/1509, 04/1112, 04/1931, 05/1079, 05/1052, 06/1213, 07/0314, and 09/02647, Red INMA G03/176, Miguel Servet CP11/0178, and 13/1944), Spanish Ministry of Science and Innovation (SAF2008-00357), the European Commission (ENGAGE project and grant agreement HEALTH-F4-2007-201413, FP7-ENV-2011 cod 282957 and HEALTH.2010.2.4.5-1), Fundacio La Marato de TV3, Generalitat de Catalunya-CIRIT 1999SGR 00241, Conselleria de Sanitat Generalitat Valenciana, and Fundacion Roger Torne. The Saguenay Youth Study is funded by the Canadian Institutes of Health Research (TP, ZP), Heart and Stroke Foundation of Quebec (ZP), and the Canadian Foundation for Innovation (ZP). The first phase of the Generation R Study was made possible by financial support from: Erasmus Medical Centre, Rotterdam, Erasmus University Rotterdam and the Netherlands Organization for Health Research and Development (ZonMw). The present study was supported by an additional grant from the Netherlands Organization for Health Research and Development (grant No. 2100.0074). EGCUT received funding through – targeted financing from Estonian Government SF0180142s08, Estonian Research Roadmap through Estonian Ministry of Education and Research, Center of Excellence in Genomics (EXCEGEN) and University of Tartu (SP1GVARENG), and from Estonian Research Council IUT2-2 grant and European Regional Development Fund. ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were

performed at the Wellcome Trust Clinical Research Facility in Edinburgh. We would like to acknowledge the invaluable contributions of Lorraine Anderson and the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney.

References

1. Martínez-Fundichely A, Casillas S, Egea R, Ramia M, Barbadilla A, Pantano L, Puig M, Caceres M, Invfest, a database integrating information of polymorphic inversions in the human genome. *Nucleic acids research* 2013; 42:D1027- D1032.

2. Cáceres A, González, JR, Following the footsteps of inversions on SNP data: From detection to association tests. *Nucleic acids research* 2015; DOI: 10.1093/nar/gkv073.

3. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al., Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008; 453:56–64.

4. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE, Characterization of six human disease-associated inversion polymorphisms. *Human molecular genetics* 2009; 18:2555–2566.

5. Magoulas PL, El-Hattab AW, Chromosome 15q24 microdeletion syndrome. *Orphanet journal of rare diseases* 2012; 7:2.

6. Mefford HC, Rosenfeld JA, Shur N, Slavotinek AM, Cox VA, Hennekam RC, Firth HV, Willatt L, Wheeler P, Morrow EM, et al., Further clinical and molecular delineation of the 15q24 microdeletion syndrome. *Journal of medical genetics* 2012; 49:110–118.

7. Roetzer KM, Schwarzbraun T, Obenauf AC, Hauser E, Speicher MR, Further evidence for the pathogenicity of 15q24 microduplications distal to the minimal critical regions. *American Journal of Medical Genetics Part A* 2010; 152:3173–3178.16.

8. Caceres A, Sindi SS, Raphael BJ, Caceres M, Gonz´alez JR, Identification of polymorphic inversions from genotypes. *BMC bioinformatics* 2012; 13:28.

9. Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, et al., The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome research* 2012; 22:1144–1153.

10. Paterson T, Law A, Arkmap: integrating genomic maps across species and data sources. *BMC bioinformatics* 2013; 14:246.

11. de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, Ophoff RA, Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC genomics* 2012; 13:458.

12. Gonzalez JR, Caceres A, Esko T, Cuscó I, Puig M, Esnaola M, Reina J, Siroux V, Bouzigon E, Nadif R, et al., A common 16p11. 2 inversion underlies the joint susceptibility to asthma and obesity. *The American Journal of Human Genetics* 2014; 94:361–372.

13. Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L et al., Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature genetics* 2012; 44:872–880.

14. Cusco I, Corominas R, Bayes M, Flores R, Rivera-Brugues N, Campuzano V, Perez-Jurado LA, Copy number variation at the 7q11. 23 segmental duplications is a susceptibility factor for the williams-beuren syndrome deletion. *Genome research* 2008; 18:683–694.

15. Zielinska DF, Gnad F, Wi´sniewski JR, Mann M, Precision mapping of an in vivo n-glycoproteome reveals rigid topological and sequence constraints. *Cell* 2010; 141:897–907.

16. Bernon C, Carre Y, Kuokkanen E, Slomianny MC, Mir AM, Krzewinski F, Cacan R, Heikinheimo P, Morelle W, Michalski JC, et al., Overexpression of MAN2C1 leads to protein underglycosylation and upregulation of endoplasmic reticulum-associated degradation pathway. *Glycobiology* 2011; 21:363–375.

17. Scott H, Panin VM, The role of protein N-glycosylation in neural transmission. *Glycobiology* 2014; 24:407–417.

18. Freeze HH, Eklund EA, Ng BG, Patterson MC, Neurology of inherited glycosylation disorders. *The Lancet Neurology* 2012; 11:453–466.

19. Xiang Z, Jiang D, Liu Y, Zhang L, Zhu L, hMAN2C1 transgene promotes tumor progress in mice. *Transgenic research* 2010; 19:67–75.

20. Wang L, Suzuki T, Dual functions for cytosolic α-mannosidase (MAN2C1) its down-regulation causes mitochondria-dependent apoptosis independently of its α-mannosidase activity. *Journal of Biological Chemistry* 2013; 288:11887–11896.

21. Yehuda R, Cai G, Golier JA, Sarapas C, Galea S, Ising M, Rein T, Schmeidler J, Muller-Myhsok B, Holsboer F, et al., Gene expression patterns associated with posttraumatic stress disorder following exposure to the world trade center attacks. *Biological Psychiatry* 2009; 66:708–711.

22. Uddin M, Galea S, Chang SC, Aiello AE, Wildman DE, de los Santos R, Koenen KC, Gene expression and methylation signatures of MAN2C1 are associated with ptsd. *Disease markers* 2011; 30:111–121.

23. Mkrtchyan G, Boyadzhyan A, Avetyan D, Sukiasyan S, Involvement of anomalous apoptosis in impairments to synaptic plasticity in post-traumatic stress disorder. *Neuroscience and Behavioral Physiology* 2014; 44:442–446.

24. Narayanan U, Ospina JK, Frey MR, Hebert MD, Matera AG, Smn, the spinal muscular atrophy protein, forms a pre-import snrnp complex with snurportin1 and importin β. *Human molecular genetics* 2002; 11:1785–1795.

25. Benyamin B, Pourcain B, Davis O, Davies G, Hansell N, Brion MJ, Kirkpatrick R, Cents R, Frani´c S, Miller M, et al., Childhood intelligence is heritable, highly polygenic and associated with fnbp1l. *Molecular psychiatry* 2013; 19:253-258.

26. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, et al., A survey of genetic human cortical gene expression. *Nature genetics* 2007; 39:1494–1499.

27. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al., Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS* Genet, 2010; 13:e1000952.

28. Guxens M, Ballester F, Espada M, Fern´andez MF, Grimalt JO, Ibarluzea J, Olea N, Rebagliato M, Tard´on A, Torrent M, et al., Cohort profile: the inma—infancia y medio ambiente—(environment and childhood) project. *International journal of epidemiology* 2012; 41:930–940.

29. Jaddoe VW, van Duijn CM, van der Heijden AJ, Mackenbach JP, Moll HA, Steegers EA, Tiemeier H, Uitterlinden AG, Verhulst FC, Hofman A, The generation r study: design and cohort update 2010; *European journal of epidemiology* 2010; 25:823–841.

30. Pausova Z, Paus T, Abrahamowicz M, Almerigi J, Arbour N, Bernard M, Gaudet D, Hanzalek P, Hamet P, Evans AC, et al., Genes, maternal smoking, and the

offspring brain and body during adolescence: design of the Saguenay youth study. *Human brain mapping* 2007; 28:502–518.

31. Johnson W, Bouchard Jr TJ, Krueger RF, McGue M, Gottesman II, Just one g consistent results from three test batteries. *Intelligence* 2004; 32:95–107.

32. Haworth C, Wright M, Luciano M, Martin N, De Geus E, Van Beijsterveldt C, Bartels M, Posthuma D, Boomsma D, Davis O, et al., The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular psychiatry 2009*; 15:1112–1120.

Figure legends

Figure 1. Inversion haplotype structure of 15q24

A) Genomic location of the region between 15q24.1-15q24.3 prone to microdeletion syndrome. **B)** Segmental duplication blocks A to E are indicated in grey. Red blocks show the possible breakpoints of microdeletion cases, adapted from Melfford et al. [6]. In particular, the thick red block is the case with the minimal deletion segment (MDS in the figure). Also, the green block illustrates the inversion discussed by Antonacci et al. [4]. coinciding with the smallest region of overlap (SRO) identified by Magoulas and El-Hattab [5]. **C)** Histogram of the multidimensional scaling analysis for the SNPs of CEU populations within the regions flanked by segmental duplications. The region C-D shows a clear clustering indicating the presence of an inversion-induced haplotype substructure. **D)** inveRsion scan over the region. A clear signal of LD differences between SNP blocks is detectable in the C-D block, suggesting the presence of an inversion polymorphism. Other signals between D-E do not present clear haplotype substructures as shown in figure C.

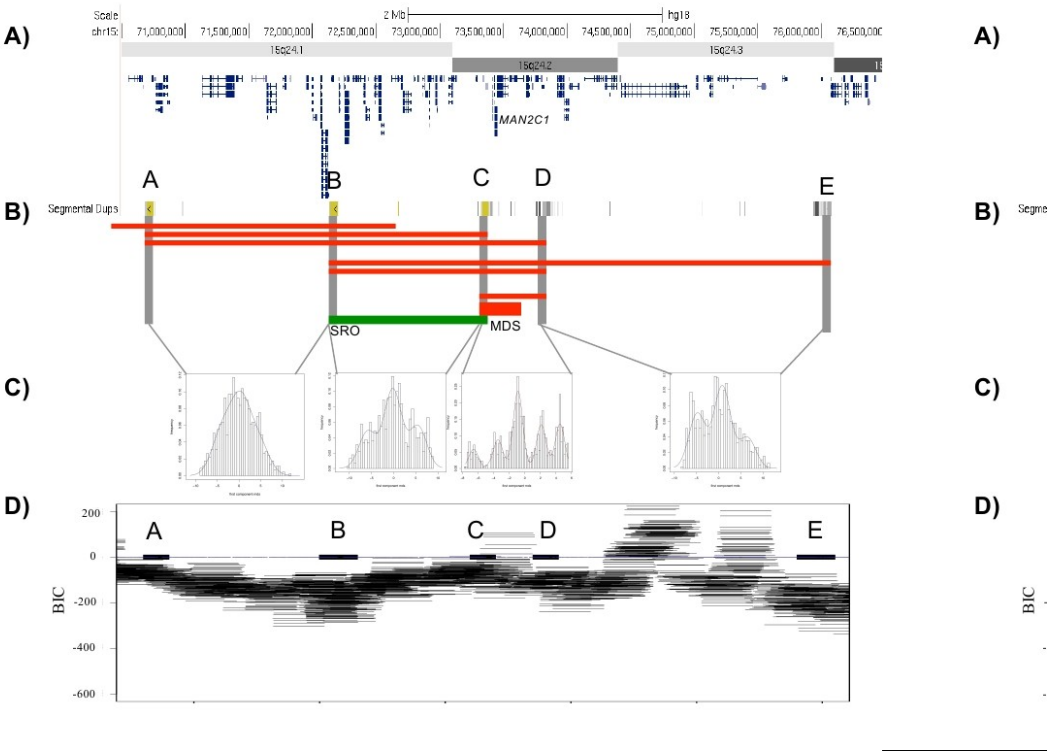
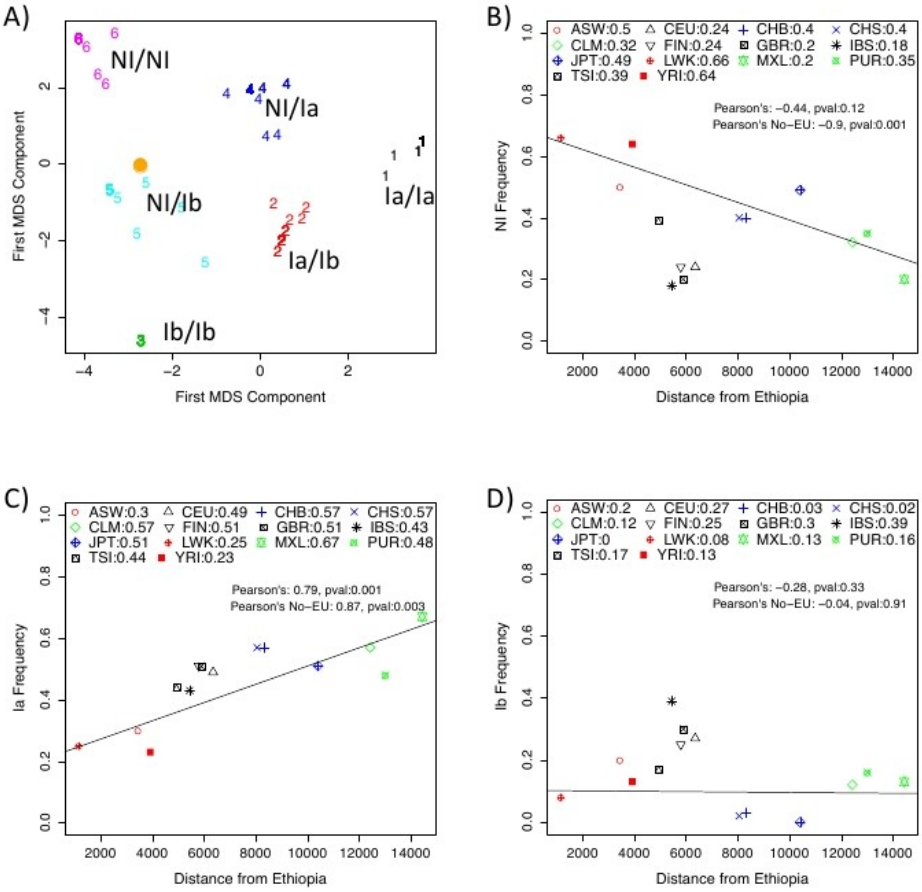


Figure 2. Global frequencies of three allele related haplotypes in 15q24.2.

A) Optimized MDS analysis with 26 SNP genotypes within the C-D region, for the CEU population. Six clusters are found representing the genotypes of three different alleles Ia, Ib, and NI. Groups 1, 2, 3, 4, 5 and 6 clusters correspond to genotypes: Ia/Ia, Ia/Ib, Ib/Ib, NI/Ia, NI/Ib, NI/NI respectively. The orange subject is a hypothetical ancestral homozygous showing that the predicted ancestral allele is related to both NI and Ib. **B)** Global uneven distribution of NI. The European frequency is lower than expected, as measured by the change in Pearson's correlation with and without Europeans. **C)** Global uneven distribution of Ia. The European group shows expected frequencies that follow an Out of Africa Expansion. **D)** Even distributions of Ib allele, excluding the European group. This group shows an increased frequency with respect to other continental groups.



A) Expression of *MAN2C1* in the peripheral blood of 882 Estonians as function of the genotype-haplotypes. **B)** Expression of *MAN2C1* in the cortex of 193 deceased individuals with no dementia. The expression of the gene was up-regulated per NI allele in blood and cortex.

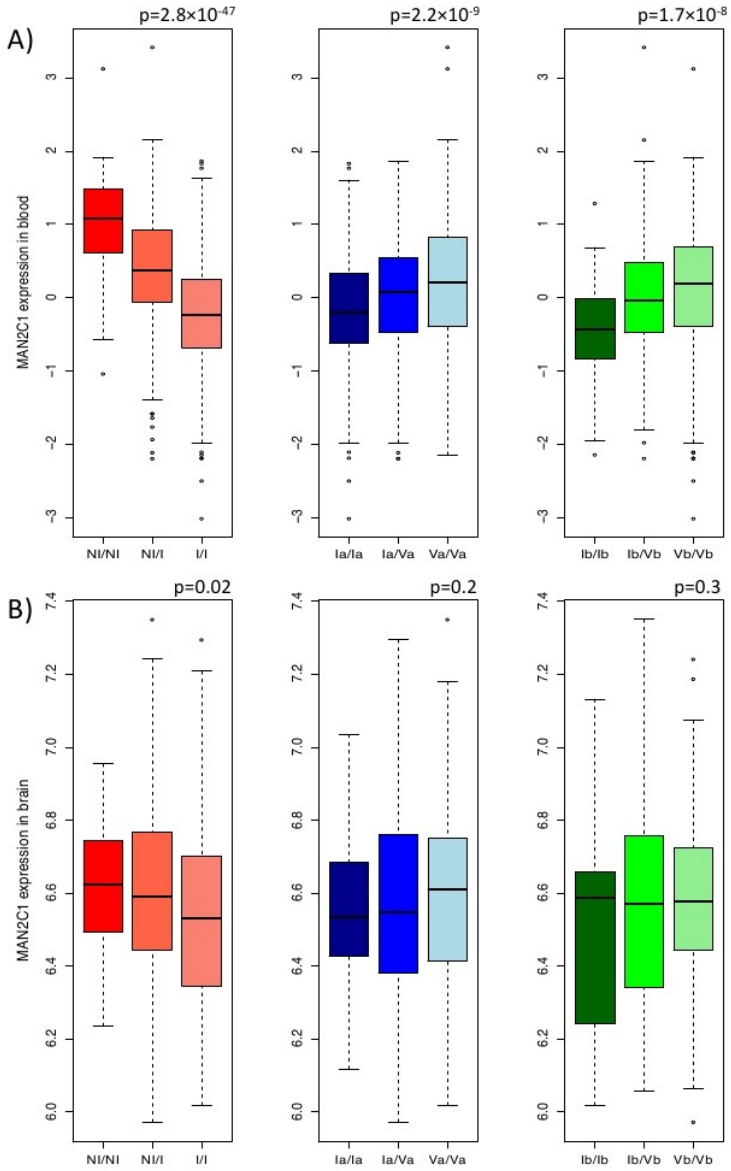


Figure 4. Association between NI, Ia and Ib alleles and the gene expressions of *SUPN* in blood and brain.

A) Expression of *SNUPN* in the peripheral blood of 882 Estonians as function of genotype-haplotypes. **B)** Expression of *SNUPN* in the cortex of 193 deceased individuals with no dementia. The expression of the gene was down-regulated per Ib allele in blood and cortex.

Figure 5. Significant deregulation of *MAN2C1* expression in the brain for NI, Ia and Ib homozygous.

MAN2C1 is ubiquitously over-expressed for NI homozygous in brain, including pons, cerebellum, frontal cortex, temporal cortex and putamen, but not in occipital cortex, where Ib homozygous highly under-express the gene. Significant results in pons, cerebellum (CRBLM), frontal cortex (FCTX), temporal cortex (TCTX) were obtained from the BRAINeQTL study while those for putamen and occipital cortex where obtained from the BRAINEAC project. Other brain regions in both data-sets showed non-significant associations.

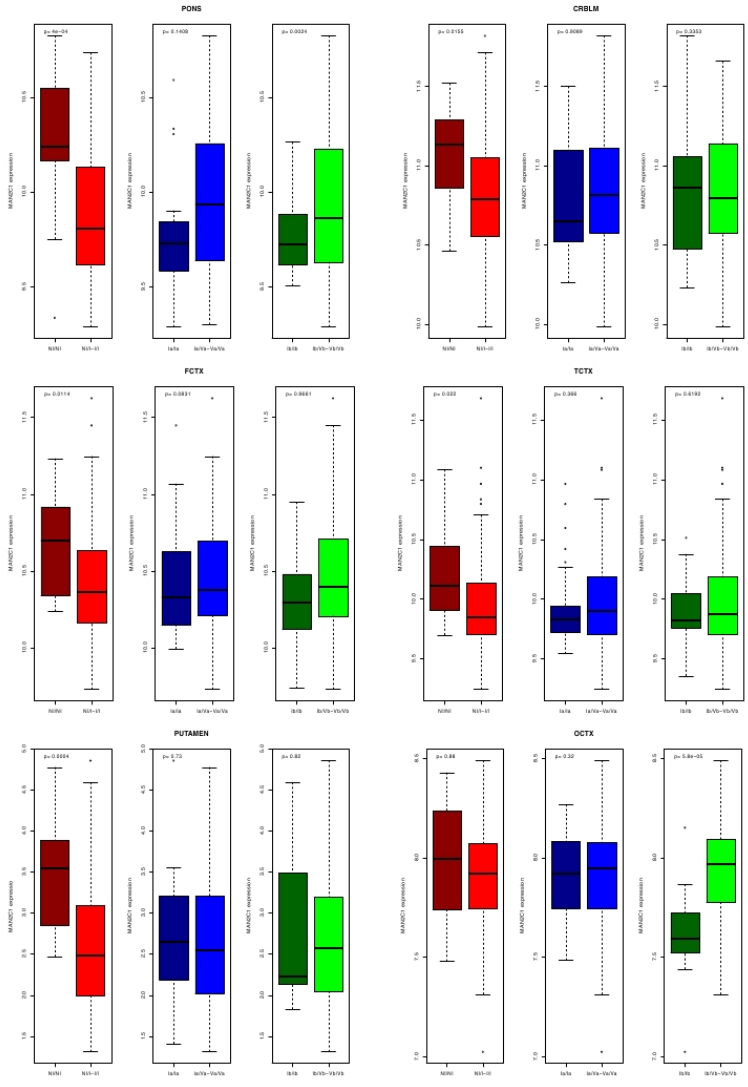
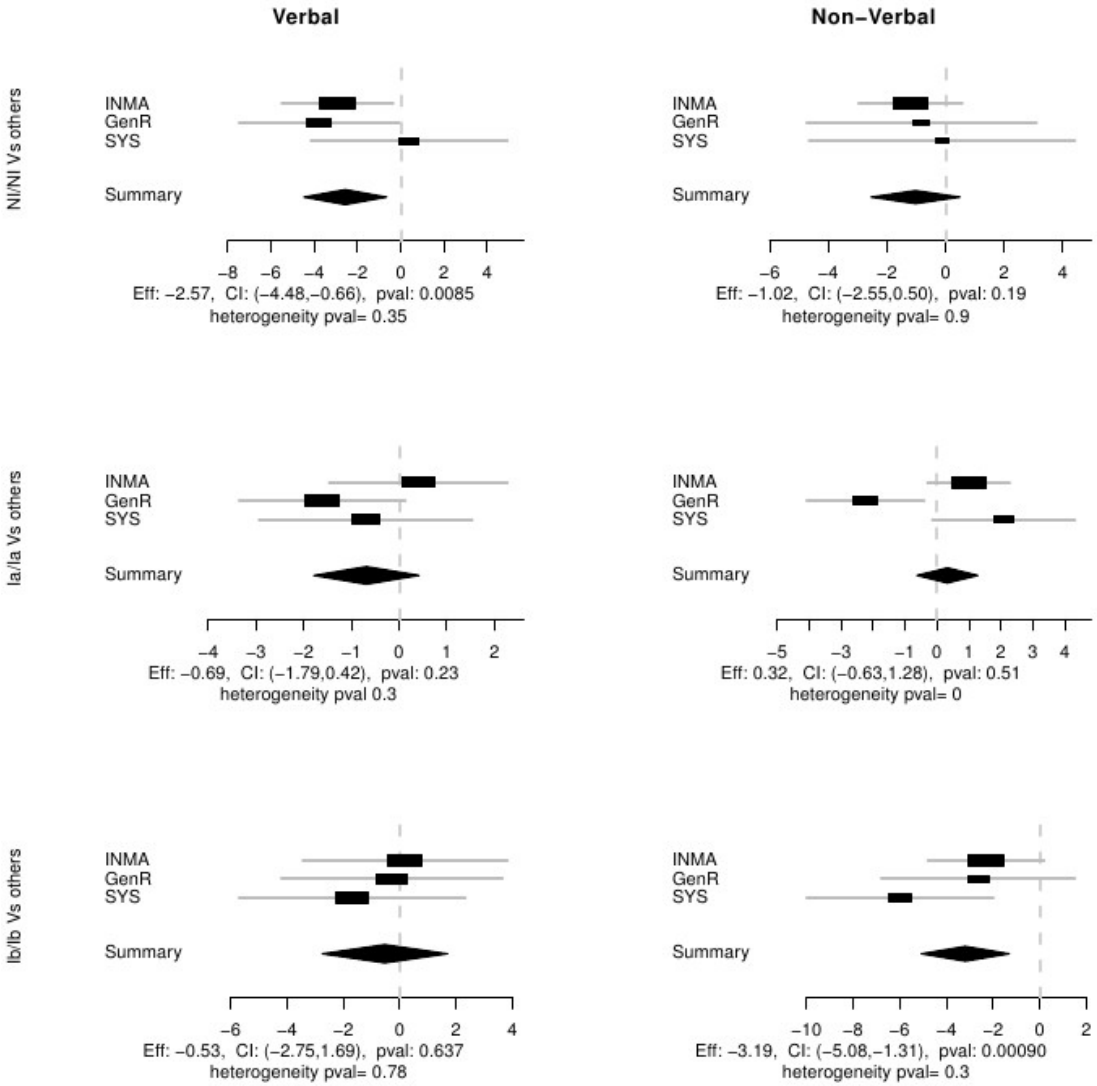


Figure 6. Association between IQ and homozygosity for NI, Ia and Ib

The figure shows the meta-analysis of three studies (INMA, SYS and GenR) for the association between IQ measures and homozygosity for the NI, Ia and Ib alleles.



Article 5.2
Publicat amb títol i contingut modificat:

Cáceres A, Esko T, Pappa I, Gutiérrez A, Lopez-Espinosa MJ, Llop S, Bustamante M, Tiemeier H, Metspalu A, Joshi PK, Wilsonx JF, Reina-Castillón J, Shin J, Pausova Z, Paus T, Sunyer J, Pérez-Jurado LA, González JR. Ancient Haplotypes at the 15q24.2 Microdeletion Region Are Linked to Brain Expression of MAN2C1 and Children's Intelligence. PLoS One. 2016 Jun 29;11(6):e0157739. doi:10.1371/journal.pone.0157739.

5.3 Rheumatoid Arthritis in North Indian population

In colaboration with B.K. Thelma laboratory, we recieved a data of 990 probands affected of Rheumatoid Arthritis (RA) of North Indian population, with 1078 controls. We performed an exploratoration of the CNV and Mosaic events. Additionally, using the available inversion prediction tools (InvClust and InveRsion), we perform a case-control study with the predicted putative inversions.

Armand Gutiérrez-Arumi, Garima Yuyal, Alejandro Cáceres, Juan Ramon González, Thelma BK, Luis Pérez-Jurado
Novel chromosomal inversions and regions of homozygosity associated with Rheumatoid Arthritis in the North Indian population.
(in preparation)

Abstract

Rheumatoid arthritis (RA) is an autoimmune disease characterized by persistent synovitis, systemic inflammation and autoantibodies. Genetic factors account for 50 % of the risk to develop RA, and heritability was estimated to be about ~60%. The largest contribution was major histocompatibility complex with a recurrent risk for sibilings ~1.8, followed by few minor variants. Nevertheless, most proportion of heritability is still unknown.

Our findings show no clear association of disorders with mosaic events, neither no clear CNV was more prevalent in cases versus controls. The analysis of overlapping regions of homozigosity returned one significant region on 4q32.1 (OR 9.32 95% CI 2.66-63.45 p-value 0.00015), from available SNPs we got a rs1158264 (OR 1.2 p-value 0.04). Three putative common inversions shows significant correlation with the disease: chr11p11.2 (OR 1.38 of non-inverted allele), chr19q13.33 (OR 1.33 of non-inverted allele), chr2q13 (OR 1.24 of inverted allele). The inversions on chr19 and chr11 affects the expression of two proteins of the same family SPIB and SPI1, highly suggestive of being involved in RA pathogenesis.

Introduction

Rheumatoid arthritis (RA) is an autoimmune disease characterized by persistent synovitis, systemic inflammation and autoantibodies. Genetic factors account for 50 % of the risk to develop RA, and in industrialized countries, RA affects 0.5-1.0% of adults, with 5-50 per 100,000 new cases annually. It is more prevalent in woman in childbearing age. The genetic component has long been established, and twin analysis shows the heritability to be about 60% [1]. The genetic basis of RA are complex, with the largest contribution of major histocompatibility complex (recurrent risk for siblings ~1.8). It has also been found a variant of PTPN22 locus outside the MHC, with odds ratios of 3.0 in homozygous variants. Despite these findings, there is still a considerable proportion of heritability is still unknown.

Methods

Dataset

We received genotype data (Illumina Human660W Quad BeadChip genotyping platform) from a total of 2179 samples of the North Indian population. We discarded 87 samples due to patterns suggestive of sample contamination and 25 identified as repeated samples. The final dataset consisted of 2068 samples, 990 labeled as probands with Rheumatoid Arthritis and 1078 as controls. Population stratification was performed with available SNPs using PCA method [2].

Mosaicism detection

For detection of copy number and copy neutral chromosomal alterations present in mosaicism we used the mosaic alteration detection (MAD) algorithm, implemented in R Genomic Alteration Detection Analysis (R-GADA) [7]. We used the B-allele frequency (BAF) measurement, derived from the ratio of probe values relative to the locations of the estimated genotype-specific clusters, for initial segmentation and detection of allelic imbalances with MAD. The $\log_2$ relative probe intensity ratio (LRR), which provides data on copy number, was then used to classify each event with abnormal BAF as copy altering (gain or loss) or neutral (reciprocal gain and loss resulting in loss of heterozygosity, LOH). Mosaic proportions were required to deviate from levels expected from constitutional (non-mosaic) changes in order to exclude homozygous

chromosomal segments inherited identical by descent and non-mosaic instances of trisomy, monosomy and uniparental disomy.

The MAD algorithm uses several parameters in order to refine the calling procedure. The parameter T controls the False Discovery Rate (FDR), and the MinSegLen indicates the number of consecutive probes that have a Bdeviation different from 0. The bigger the false discovery rate, more probable is to get false positives. We run MAD with three parameters, T=4 and minSeg=25, T=8 and minSeg=25, and T=9 and minSeg=75. We then filtered out several false positives, and we performed a visual inspection of the plots for curation of the detected events.

Copy Number Variation

For CNV calling, we used the *PennCNV* software. A brief summary of the filtering steps was as shown:

	Total CNV counts	**False Positive Rate**
Raw PennCNV	260,747	-
PennCNV 10-SNP	194,303	0.9
Outlier-filter	9,563	0.05* (of CNVs bigger than 100 SNPs)

Table 1. Total CNV counts with an estimation of False Positive Rate manually checking 100 CNVs at random .

We first applied a correction of the log ratio obtained from raw data based on GC content. After this normalization, *pennCNV* analysis yielded a total of 260,747 CNV events.

A quick test to estimate false discovery rate revealed a small subset of samples (n=172, 7.8%) with a very high number of CNVs. Visual inspection of chromosomal plots of these samples revealed that most (or all) CNV calls in these samples were false positives due to technical artifacts or normalization problems. Therefore, we applied an outlier-filter to discard samples with excessive CNV calls, resulting in a total of 9,563 CNVs in the entire dataset. We repeated the test of false discovery rate on those events with CNVs bigger than 100 SNPs. To assess for additional quality information of the CNVs, *parseCNV* was used. We finally applied a manual curation and visual inspection of the apparently common CNVs.

A case-control association analysis with the CNV calls was done with the PLINK package. P-values were corrected after 1000 permutations (adaptive permutations).

Overlapping Regions of Homozygosity (ORH)

Considering the significant inbreeding of the Indian population, it is likely that homozygosity for recessive alleles in regions identical by descent can also contribute to the susceptibility to this multifactorial disorder especially in this population. We ran

PLINK on *PennCNV* output files to define regions of homozygosity shared by multiple samples [3]. A total of 1453 overlapping regions were identified.

We randomly selected 40 of them for visual inspection of the chromosomal plots and detected no false positives. Then, we performed a case control association study. ORH with positive signals were also inspected and manually curated. We then defined whether these ORH overrepresented in patients shared identical genotypes and haplotypes that could be identical by descent (IBD). We then reran a targeted association test in all samples with the overrepresented SNPs in ORH.

Inversion Calling

In order to identify potential genomic inversions in this dataset located at intervals previously defined in other populations, we applied two different algorithms on the available SNP data [4]. The first algorithm, inveRsion, is based on differences in LD between SNP blocks across predicted inversion breakpoints. This algorithm searches for possible breakpoints, based on differences of Bayes Information Criterion (BIC). When BIC is greater than zero, it suggests that the chromosomes of some individuals are more likely than not to harbor an inversion in between the tested interval. In a genome-wide search in 180 CEU individuals from HapMap using 0.4 Mb window sizes, we previously identified 174 putative inversion signals with inveRsion [5][6]. The second algorithm, invClust, is based on *PFIDO* [2] and uses multivariate analysis of SNPs within a region to predict inversion-related genotypes [8]. As *invClust* needs the a priori definition of breakpoints of the region to scan, we performed a targeted analysis of the regions detected by inveRsion. Then *invClust*, automatically applies a clustering k-means method to identify the main genotype groups (clusters) in the dataset. Theoretically, an inversion will detect two main haplotypes (I and N alleles), identifying three groups or main clusters(homozygous for the inverted and non-inverted alleles and the heterozygous ones). To infer the error rate of inversion calling for each of the 174 inversions, we used a dataset of 2200 European trios and checked the Mendelian transmissions of predicted inversion haplotypes. We discarded regions with high error rate, suggestive of poor calling. A total of 49 regions with well-clustered haplotypes

suggestive of putative inversions were found in the North Indian datasets. We also studied functional consequences of the putative inversions using blood transcriptome data from the same 180 HapMap individuals. In 24 of the 49 regions, there was a strong correlation of the predicted inversion haplotypes with expression of genes within the interval.

We then used the *invClust* program to define the haplotypes for the 49 predicted inversions in the North Indian dataset and performed a case-control association test.

Statistical Analyses

In all the case-control studies with CNVs and inversions, p-values were corrected by multiple-testing using Bonferroni correction, applying the appropriate threshold. Since we tested 24 well-clustered inversions in this population, a threshold < 0.00208 was considered significant.

For the case-control analyses with ORHs, we applied a chi-square test to each ORH identified, using the PLINK package. For the first ORH analysis, we also applied Bonferroni correction and defined the level of significance bellow a p-value of 0.000168 (0.05/299). Other p-values were not initially corrected, and are individually commented.

Results

Mosaic events

Summary of results (sample ID, type of rearrangement, chromosome arm involved, coordinates of the initial and final probe within the rearrangement and affection status of the carrier (case or control).

ID	type	% of cells	chr	IniProbe	EndProbe	Individual
GWACO571	monosomy	90,22	7	1	159122532	Control
GWACO571	trisomy	19,52	8	1	146271129	Control
HFC1332	trisomy	37,61	8	1	146271129	Control
HFC1332	trisomy	38,12	9	1	141122599	Control
HFC1332	gain 1q	38,19	1q	145394955	249210707	Control
HFC1332	gain 19q	37,98	19q	1	25000000	Control
GWACO229	monosomy	70,74	7	10704	159122532	Control
HFRAP179	UPD	80,30	9p	1	26524684	Case
HFRAP285	UPD	24,65	9q	74879449	141122599	Case
RA672	UPD	24,62	2q	220581322	243101834	Case
MRA1125	Deletion	67,93	17q	43651233	43665677	Case
RA161	Deletion	39,90	11p	2479195	3131741	Case
RA748	Deletion	31,75	20q	31619922	49120395	Case
RACO1070	Deletion	43,54	16q	87564964	89066527	Control

Table 2 Mosaic events found, with the location, and the percentage of cells affected.

We detected whole chromosome rearrangements and/or multiple events in three control individuals, suggestive of clonal proliferation of a hematological tumor. GWACO571 (monosomy 7 and trisomy 8), HFC1332 (trisomy 8, trisomy 9, and trisomies 1q and 19p likely due to translocation 1q;19p), GWACO229 (monosomy 7).

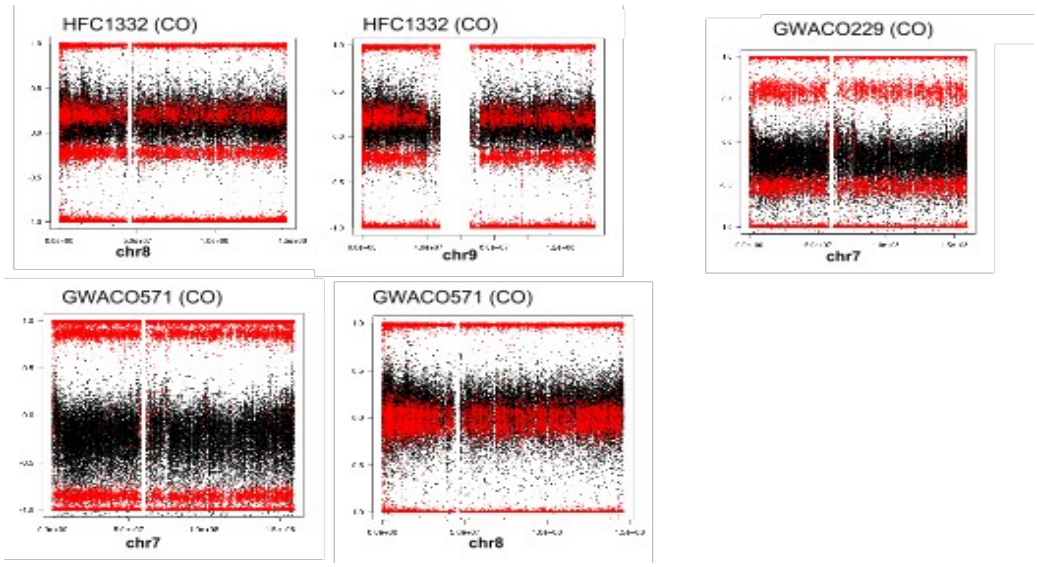


Fig 1. Images of the mosaicism found on chr7, chr8, chr9. The control sample with a monosomy on chr7 it is likely to have a leuckemia. Molecular karyotyping idiograms of the chromosomes with detected segmental rearrangements. LRR values represented as black dots (scale on the left) and BAF values represented as red dots (scale in the right).

If we do not consider the mentioned three control individuals with whole chromosome rearrangements, the number of individuals with segmental mosaicism was slightly enriched in RA patients (n=6, 3 with segmental deletions and 3 with segmental UPDs) with respect to controls (a single individual with a mosaic segmental deletion). However, each detected event was in a different genomic region with no recurrence. The information about the age of the individuals with mosaic events is still missing.

CNVs

After discarding samples with mosaic events and those with >= 25 CNV calls per sample, pennCNV detected a total of 9,563 CNV events in the final dataset of 2068 samples (990 RA cases and 1078 controls).

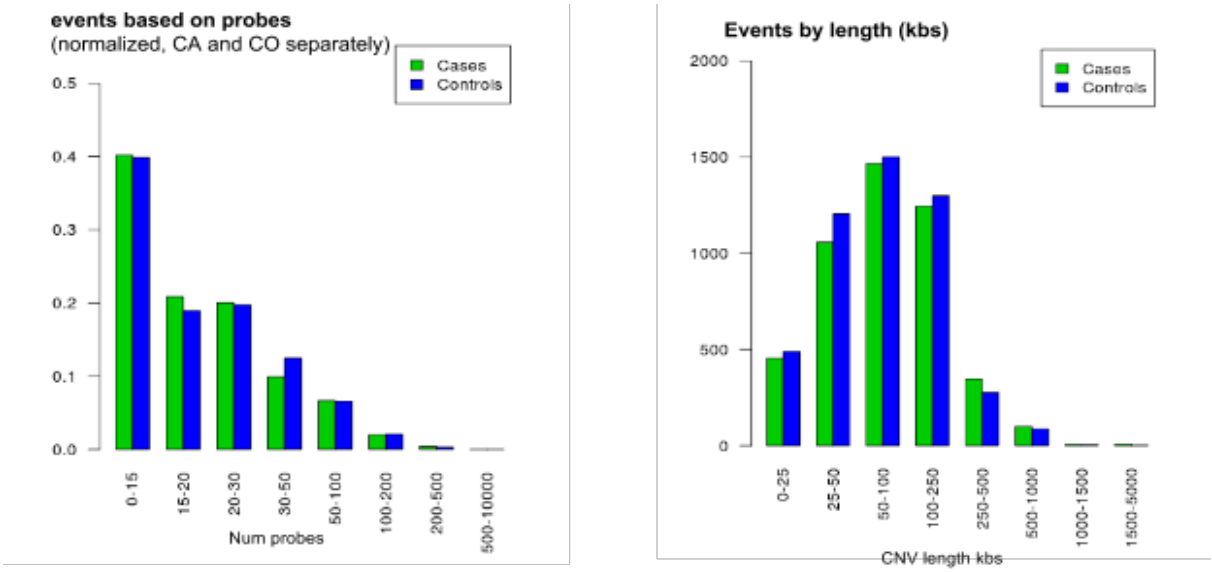


Fig 2. After normalitzation number of CNV events found. The first graph shows a distribution of the percentage of events with the number of probes. The second one, shows the number of CNV events considering its length.

After normalization, cases and control samples harbored a similar amount of CNV events. Large CNVs (>100 probes), 216/9563 (2.2%), were manually curated. A total of 7 were discarded, mostly due to the fact that they corresponded to pericentromeric artifacts. Smaller CNVs (<100 probes) 9350 calling events with no additional filtering

CNV case-control study

Only two CNVs on chromosomes 22 and 8, respectively, were found significantly overrepresented in cases with respect to controls after multiple test correction, and a third on chromosome 7 barely reached significance.

Coordinates	Event	Cases	Controls	p-value	Quality
8:43658198-43910848	DEL	77	21	0.000999	fail
22:21064159-21066291	DUP	39	14	0.007992	fail
7:61490330-62075724	DEL	25	6	0.016983	fail

Table 2. Significant CNV and its region that fails posterior quality filters.

However, despite the positive signals with *parseCNV*, visual inspection of the plots suggested that these CNV calls could also be artifacts. All three putative CNVs were located in regions enriched in segmental duplications and other repeats. We then analyzed the LRR value distribution in the entire dataset. The distribution of LRR

values for all samples at these loci showed a monomodal curve with large tails based on standard deviation, instead of the expected three-several peak distribution in the case of alleles with different copy number (data not shown).

Runs of Homozygosity (ROH)

The case-control association test revealed a single region located on *chr4q32.1* that was found in homozygosity region significantly more enriched in cases than in controls. The minimal interval of overlap corresponded to the chr4 coordinates 160244776-161281594.

Region coordinates	Ca	Co	OR	95%CI	p-value
chr4:160244776-161281594	18/990	2/1078	9.32	2.66-63.45	0.00015

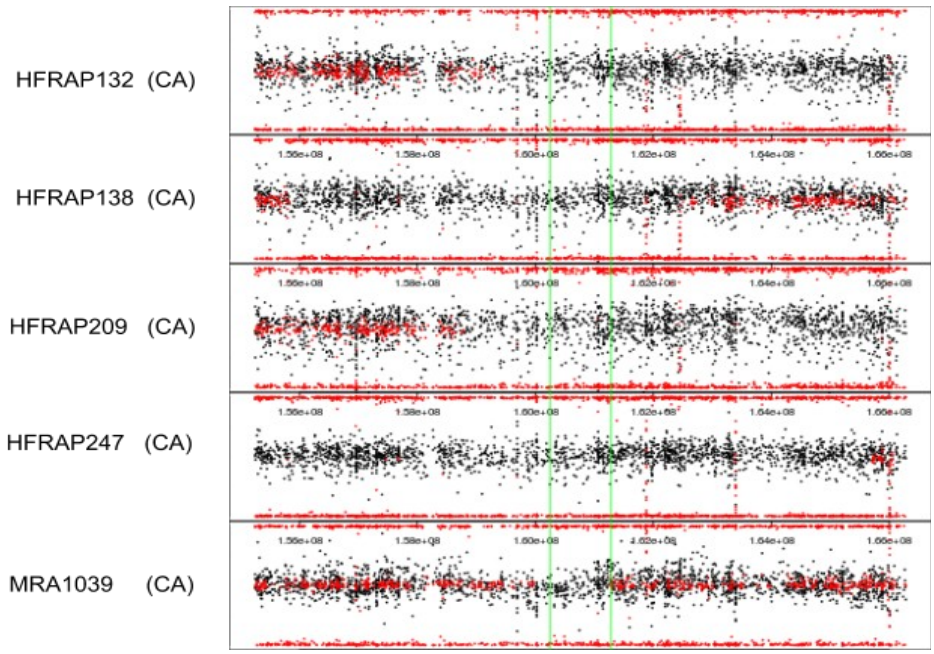


Fig 3. Homozygous overlapping region on chr4p32.1 between 5 of the 18 representative samples. Region of homozigosity significant overlapping 18 cases against 2 controls. Probands with regional homozygosity: HFRAP132, HFRAP138, HFRAP209, HFRAP247, HFRAP247R, MRA1039, MRA1049, MRA1066, MRA1066D, MRA1133, RA045, RA056, RA059, RA219, RA558, RA588,RA715, RA718 Controls with regional homozygosity: RACO1188, RACO1215

Regional plots of 5 representative individuals with regional homozygosity, all of them RA cases, as it is shown in Fig 3. The minimal region of overlap is shown in between green lanes.

Genomic features of the ORH on chromosome 4q32.1. This region contains a single coding gene *RAPGEF2*. *RAPGEF2* has already being identified as a susceptibility gene for RA and other autoimmune diseases (supplementary information with refs).

We further searched for potential recessive allele in the regions. A total of 115 SNP with MAF > 0.01 were identified in the ORH. We then rerun a case-control study with those SNPs. None of the SNPs was significant using Bonferroni correction with a threshold of 0.05/25 = 0.002 (the p-values are not corrected).

estimate	lower	upper	p-value_chi	Snp.name	Allele value	Num CA	Num CO	Total CA	Total CO
1.20	1.0006	1.44994	0.04907	rs1158264	CC	693	711	990	1078

Snp.name	Allele value	Num CA (IBD)	% num CA (IBD)	Population allele freq %	val
rs1158264	CC	13	0.81	65.95	

Table 3. From available SNPs, the putative recessive allele on the snp rs1158264.

Inversion association study

After inversion calling (49 selected regions with putative inversions and well-differentiated clustering by InvClust), we performed a case-control association analysis with the inversions. After correcting by multi-dimensional scaling of the population and by gender, 3 of the 49 putative inversions showed a genotype distribution that was significantly different in RA cases with respect to controls:

Inversion coordinates	**OR**	**p-value**	**model**
chr11:49852305-54818608	1.38	0.0003	dominant
chr19:55269524-55818494	1.33	0.0020	dominant
chr2:109820409-110858548	1.245	0.0022	additive

Table 4. From significant inversions, chr11p11.2 (OR 1.38 of non-inverted allele), chr19q13.33 (OR 1.33 of non-inverted allele), chr2q13 (OR 1.24 of inverted allele).

In order to check the homogeneity of North Indian population, a PCA was performed using available SNPs. Two main subpopulation structure. Nevertheless, the same inversion frequencies were detected in the two main clusters.

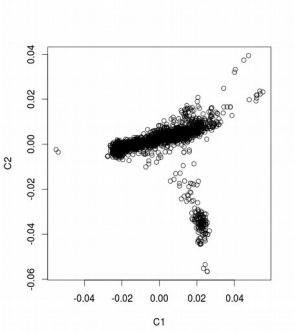


Fig 6. Two main clusters in North Indian population, shown the same frequencies in controls and cases (less than 1% of differences).

Although PCA detected two different groups, the distance in between was very low. After correcting the associations with inversions by PCA, the correlations still remained.

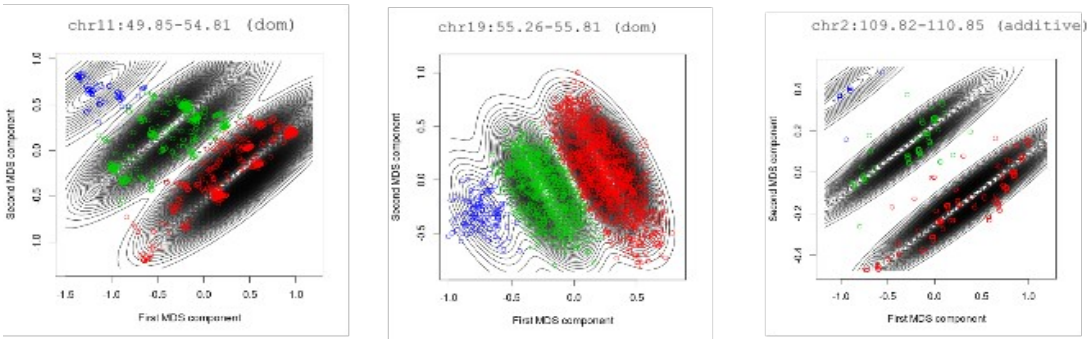


Fig 7. The three putative inversions detected. Blue represents the homozygous inverted allele, green the heterozygous whereas red the homozygous for the non-inverted allele.

As it is shown in Fig 7, from 3 putative inversions detected, 11p11.2 seems to have at least 6 main clusters instead the 3, suggesting 3 inversion allele.

Discussion

We also used a database of 2561 European trios in order to define the accuracy of the haplotype prediction by InvClust for these inversions. We detected a Mendelian error rate of 39/2561=0.015, 56/2561=0.021 and 3/2561=0.001, in the analyses of the chr11, chr19 and 2q13 inversions respectively. Therefore, the 2q13 inversion is the strongest candidate, while the chr11 and chr19 inversions would still benefit from better clustering.

As previously stated, we previously established a correlation of the predicted inversion alleles with gene expression of regional genes in blood, using the Hapmap and Estonian population datasets. The two putative inversions on chr11 and chr19 did show clear

regulatory effects on local genes that were differentially expressed depending on the predicted allele. However, no significant effect on local genes using blood samples was observed for the chr2 inversion.

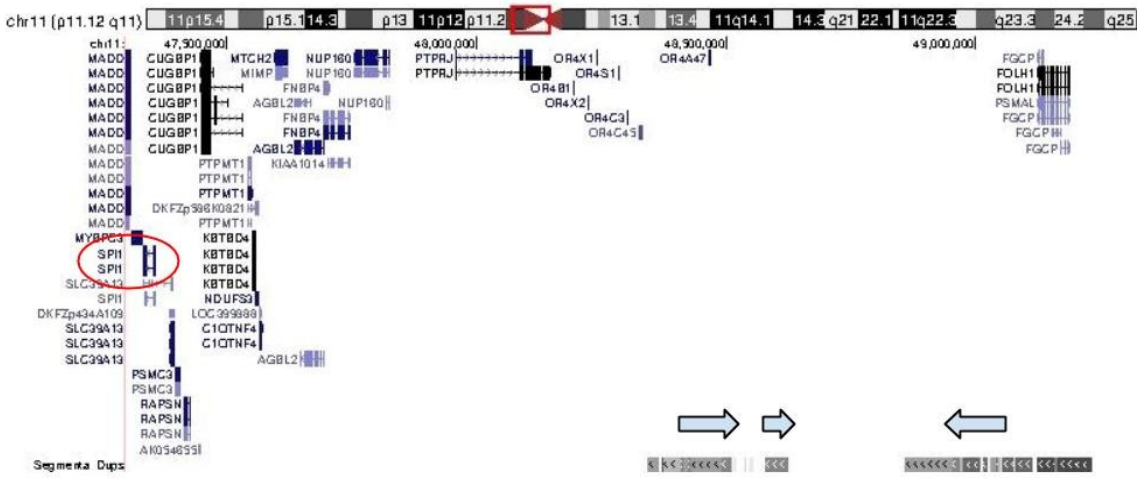


Fig 8. Blue arrows represent were the inversion is found (blue arrows are upper the segmental duplications). Grey segmental duplications means 90-98% similarity. In red circle the SPI1 were correlates with inversion genotype.

The only gene differentially expressed by chr11 inversion was *SPI1* (spleen focus forming virus (SFFV) proviral integration oncogene), which is located 2 Mb proximal to the predicted inversion breakpoint. Diseases associated with SPI1 include primary effusion lymphoma. GO annotations related to this gene include *RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity* and *sequence-specific DNA binding transcription factor activity*. An important paralog of this gene is ENSG00000142539. Their function is to bind to the PU-box, a purine-rich DNA sequence (5'-GAGGAA-3') that can act as a lymphoid-specific enhancer. This protein is a transcriptional activator that may be specifically involved in the differentiation or activation of macrophages or B-cells. Also binds RNA and may modulate pre-mRNA splicing (By similarity).

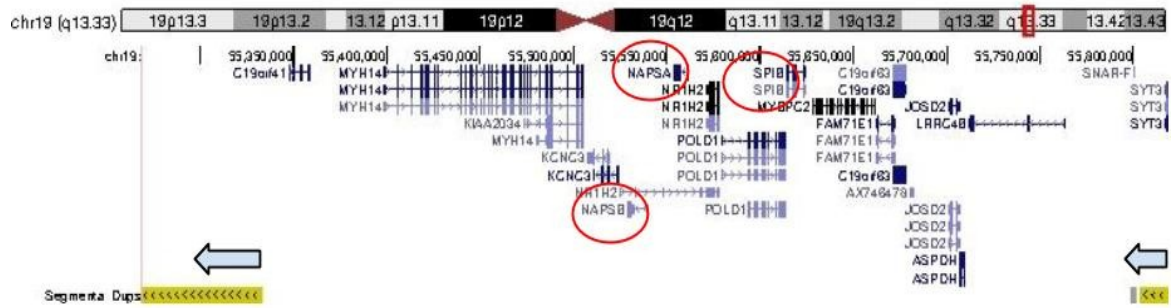


Fig 9. Blue arrows represent were the inversion is found (blue arrows are upper the segmental duplications). Yellow segmental duplications means 98-99% similarity. In red circle the SPIB were correlates with inversion genotype, as well as NAPSA, and NAPSB.

In case of the chr19 inversion, three genes and a pseudogene were differentially expressed by inversion alleles: *NAPSA* (napsin A aspartic peptidase), *NAPSB (pseudogene), PSG6 (*a member of the pregnancy-specific glycoprotein family) *& SPIB.*

NAPSA: The activation peptides of aspartic proteinases play role as inhibitors of the active site. These peptide segments, or pro-parts, are deemed important for correct folding, targeting, and control of the activation of aspartic proteinase zymogens. The biological function of NAPSA may be involved in processing of pneumocyte surfactant precursors.

PSG6: The PSG genes are a subgroup of the carcinoembryonic antigen (CEA) family of immunoglobulin-like genes, and are found in a gene cluster at 19q13.1-q13.2 telomeric to another cluster of CEA-related genes. The PSG genes are expressed by placental trophoblasts and released into the maternal circulation during pregnancy, and are thought to be essential for maintenance of normal pregnancy.

SPIB: Interestingly, the protein encoded by *SPIB* is also a transcriptional activator similar to SPI1 that binds to the PU-box (5'-GAGGAA-3') and acts as a lymphoid-specific enhancer. Four transcript variants encoding different isoforms have been found for this gene. Promotes development of plasmacytoid dendritic cells (pDCs), also known as type 2 DC precursors (pre-DC2) or natural

interferon (IFN)-producing cells. These cells have the capacity to produce large amounts of interferon and block viral replication.

Strikingly the putative inversions 19q13.33 and 11p11.12 regulates the expression of the same family of proteins SPIB and SPI1 (enhancers via PU-box), and it is highly suggestive that may have a role in rheumatoid arthritis disease.

Bibliography

[1]. Amos, C. I., Chen, W. V, Lee, a, Li, W., Kern, M., Lundsten, R., …, Gregersen, P. K. (2006). High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. Genes and Immunity, 7(4), 277–286. doi:10.1038/sj.gene.6364295

[2]. Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *PloS One*, *7*(7), e40224. doi:10.1371/journal.pone.0040224+

[3]. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–75. doi:10.1086/519795

[4]. González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., … Pérez-Jurado, L. A. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *American Journal of Human Genetics*, *94*(3), 361–72. doi:10.1016/j.ajhg.2014.01.015

[5]. Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. *BMC Bioinformatics*, *13*(1), 28. doi:10.1186/1471-2105-13-28

[6]. HapMap project [http://hapmap.ncbi.nlm.nih.gov/]

[7]. González, J. R., Rodríguez-Santiago, B., Cáceres, A., Pique-Regi, R., Rothman, N., Chanock, S. J., … Pérez-Jurado, L.A. (2011). A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics*, *12*(1), 166. doi:10.1186/1471-2105-12-166

[8]. Cáceres, A., González, J. R. (2015). IncClust. *Nucl Acid Res*, *12*(1), 166. doi:10.1186/1471-2105-12-166

[9]. Salm, M. P., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., … Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research*, *22*(6), 1144–53. doi:10.1101/gr.126037.111

5.4 Ulcerative Colitis in North Indian population

In colaboration with B.K. Thelma laboratory, we recieved a data of 806 probands affected of Rheumatoid Arthritis (RA) of North Indian population, with 1026 controls (1832). We performed an exploratoration of the CNV and Mosaic events. Additionally, using the available inversion prediction tools (InvClust and InveRsion), we perform a case-control study with the predicted putative inversions.

Armand Gutiérrez-Arumi, Garima Yuyal, Alejandro Cáceres, Juan Ramon González, Thelma BK, Luis Pérez-Jurado
7q35 and 16p11 inversions with 4q32.1 genes as susceptibility factors in Ulcerative Colitis in North Indian population.
(in preparation)

Abstract

Ulcerative Colitis (UC) is an autoimmune disease is a subcategory of inflammatory bowel disease that causes inflammation and ulcers in the colon. The prevalence it varies dramatically according sample origin. In European population is rare (less than 1%), whereas in Japanese population is highly prevalent (20%-25%). The largest contribution was observed in HLA-DRB1 region, but in any case, most of the heritability of UC (>70%) has still not been characterized.

Our findings show no clear association of disorders with mosaic events, although a rare 11p mosaic event appear in three out of four controls and only one case (the case with higher percentage). No CNV was more prevalent in cases versus controls. The analysis of overlapping regions of homozigosity returned one significant region on 3p23.3 (OR 5.36 CI 95% [1.9-19.32]). Two inversions appear to be correlated with UC, the 7q35 (OR 1.28 non-inverted allele p-value 0.021) and 16p11 (OR 1.26 inverted-allele p-value 0.0031).

Introduction

Ulcerative Colitis (UC) is an autoimmune disease is a subcategory of inflammatory bowel disease that causes inflammation and ulcers in the colon. UC have an incidence of 1.2 - 20.3 cases for every 100,000 persons per year, with a prevalence of 7.6-246 cases per 100,000 per year (Colitis, U. (2003), Danese, S., & Fiocchi, C. (2015)), with varies considerably among population. In European population is rare (less than 1%), whereas in Japanese population it is highly prevalent (20%-25%). The most reproducible association observed is HLA-DRB1, but this variant precisely it has low prevalence to Europeans (Fernando, M. M. a et al (2008)). In any case, most of the heritability of UC (>70%) has still not been characterized.

Methods

Mosaicism

Detection of clonal mosaic events was based on assessment of allelic imbalance and copy-number changes. We used the B-allele frequency (BAF) measurement, derived from the ratio of probe values relative to the locations of the estimated genotype-specific clusters, for initial segmentation using the mosaic alteration detection (MAD) algorithm implemented in R Genomic Alteration Detection Analysis (R-GADA) [7]. The BAF and log2 relative probe intensity ratio (LRR), which provides data on copy number, were used to classify each event as copy altering (gain or loss) or neutral (reciprocal gain and loss resulting in loss of heterozygosity, LOH). Mosaic proportions were required to deviate from levels expected from constitutional (non-mosaic) changes in order to exclude homozygous chromosomal segments inherited identical by descent and non-mosaic instances of trisomy, monosomy and uniparental disomy.

The MAD algorithm, uses several parameters in order to refine the calling procedure. The parameter T controls the False Discovery Rate (FDR), and the MinSegLen indicates the number of consecutive probes that have a Bdeviation different from 0. As bigger the false discovery rate, more probable is to get false positives, but we consider adequate to run it with T=4 and minSeg=25 and T=8 and minSeg=25 and T=9 and minSeg=75. From this results, we filtered out several false positives, and we checked manually the remaining results.

CNVs

For the calling of CNVs, was employed *pennCNV* software. A brief summary of the filtering steps was the following.

	Total CNV counts	**False Positive Rate**
Raw PennCNV	210,024	-
PennCNV 10-SNP	157,379	0.92
Outlier-filter	7,401	0.00* (of CNVs bigger than 100 SNPs)

Table 1. Filtering steps of CNV counts and estimating a False Positive Rate selecting 100 CNVs at random.

We applied correction of the log ratio based on GC content in raw data before launching *pennCNV*, yielding a total of 210,024 CNV events.

After perform a quick test to asses an approximate rate of false discovery rate, it became evident that a small subset of samples (7.6 %) had a very high CNV-counts, and visual inspection revealed a majority of the CNVs in these samples to be false. To eliminate the excessive CNV-count am outlier-filter was applied, ending-up with 7,401. We repeated the test of false discovery rate on those events with CNVs bigger than 100 SNPs, and in a subsample of 20, we don't get any false CNV (we could visually confirm).

Nevertheless, not all CNVs recorded are real and could be artifacts. For example, the CNVs that appear in segmental duplication zones, or those close to the centromere or at the flanking regions of the chromosomes, could be as false positive. To assess an additional quality information of the CNVs, *parseCNV* was used.

IBD

We ran *plink* on *pennCNV* output files to perform the calling of regions of low heterozygosity (Identity By Descent) [3]. An amount of 1453 overlapping regions were found after launching.

This zones are less prone to get false positives, and we randomly validated visually 40 of them with no false positives. Afterwards, we perform a case control association study, and we checked out manually those that appear to be significant.

Inversion Calling

In order to characterize the different inversions, we applied two different algorithms based on SNP data. The first algorithm, inveRsion, is based on differences in LD between SNP blocks across inversion breakpoints. This algorithm searches for possible breakpoints, based on differences of Bayes Information Criterion (BIC) which, if greater than zero, indicates that the chromosomes of some individuals between tested interval are more likely to be inverted than not. We ran inveRsion using 0.4 Mb window sizes genome-wide, for 180 CEU individuals from HapMap III [6]. The second algorithm, invClust, is based on *PFIDO* [2] and is based on the detection of inverted-related genotypes (or clusters) based on multivariate analysis of SNPs within the inverted region. As *invClust* needs the breakpoints of the region to scan, we added the regions already found by inveRsion. Then *invClust*, automatically applies a clustering k-means method to identify inversion genotypes. Ideally an inversion is formed by three clusters, that is two main haplotypes the inverted and the non-inverted allele. Therefore the three genotypes should be when both alleles of a given individual are homozygous for the inverted or non inverted allele, and the heterozygous ones. This strategy was successfully applied in a previous work [4].

With all the candidate inversion regions, we have a European trio's 2561 dataset, and we checked the Mendelian error rate. Filtering out, those that present more than 30 Mendelian errors, we got 49 putative inversions. Of them, 24 have a correlation in blood expression of some genes.

The *invClust* program was used in order to perform the calling of the inversions, and a collected dataset of inversions associated with data expression.

Population stratification

The whole set of unrelated samples shows high rates of inbreeding, with high homozygouse regions. The underlaying population structure from PCA method with available SNPs [1] shows two clear main clusters [Fig. 1].

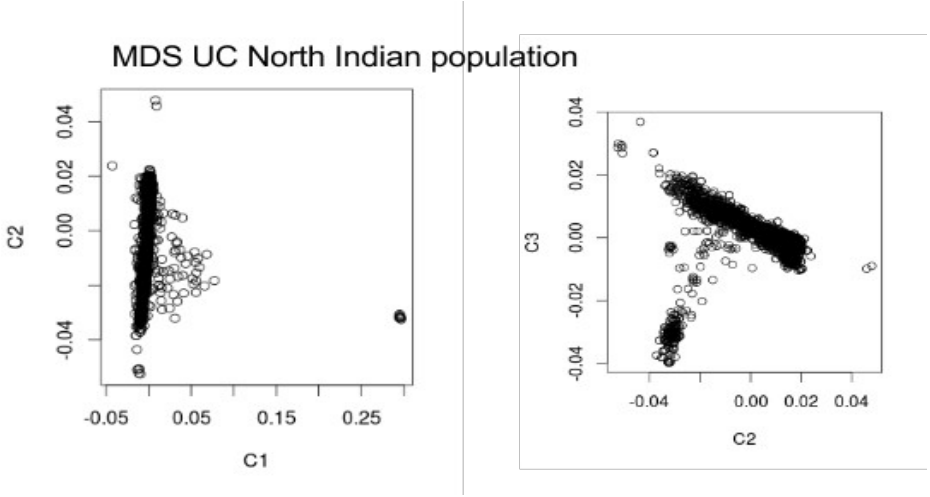


Fig. 1 The samples shows high degree of homogeneity with two main clusters of North Indian population.

No correlation of the data was related with population stratification (neither IBDs or inversion calling).

Statistical Analysis

In all the case-control statistical analysis performed, the p-values are corrected by multiple-testing using Bonferroni correction, and applying the corresponent threshold.

The CNV calling as well as the estimation of p-values, are computed using PLINK package. The showed p-values are corrected using performing 1000 permutations (adaptive permutations). So the values are genome-wide significant beyond 0.05.

For the IBD case-control analysis, a discover of the different regions of homozigosity, PLINK package were used. It describes different pools of similar overlapping samples. We selected those pools were the number of overlapping samples were maximized, and performing a case-control chi-square test on each of it. The p-values are not corrected, and in the different cases the considered p-value threshold are commented in each case. For example, for the first IBD analysis, we asses that a given value surpasses Bonferroni correction if it has less of 0.000168 (0.05/299) value.

For the inversions calling, the p-values are corrected by multiple-testing according Bonferroni correction method. In our case, we perform a test using 24 reliable

inversions in this population, so a threshold lower than 0.00208 we considered as significant.

Results

Mosaicism

No mosaic events were more prevalent in cases than in controls. In table 1 there is a summary of results (sample ID, type of rearrangement, chromosome arm involved, coordinates of the initial and final probe within the rearrangement and affection status of the carrier (case or control).

Sample ID	type	chr	IniProbe	EndProbe	BDev	% of Cells affected	
HFC1063	UPD	11	70864	47500267	0.114	22.91	1
HFC1169	UPD	11	70864	45141700	0.005	10.19	1
DMC14CA	UPD	11	70864	47153558	0.360	72.06	2
GWACO465	UPD	11	73365695	134940416	0.189	37.86	1
GWACO232	UPD	11	70864	47468569	0.104	20.97	1
GWACO372	UPD	20	31457887	62435629	0.099	19.84	1
GWACO372	UPD	7	113932460	144940416	0.031	6.26	1

Sample ID	type	chr	IniProbe	EndProbe	BDev	% of Cells affected	
GWACO571	monosomy	7	10704	159122532	0.409	90.00	1
GWACO571	trisomy	8	10213	146271129	0.055	19.82	1
HFC1332	trisomy	8	10213	146271129	0.116	37.81	1
HFC1332	trisomy	9	11428	141122599	0.1186	38.35	1
GWACO229	monosomy	7	10704	159122532	0.272	70.59	1

Table 1. Summary of all mosaic events found (mosaic segmental UPD , monosomies and trisomies).

No mosaic interstitial deletion were detected in Colitis Ulcerosa dataset.

Strikingly a rare 11p mosaic event appear in three out of four controls and only one case (the case with higher percentage) as it can be shown in fig 1.

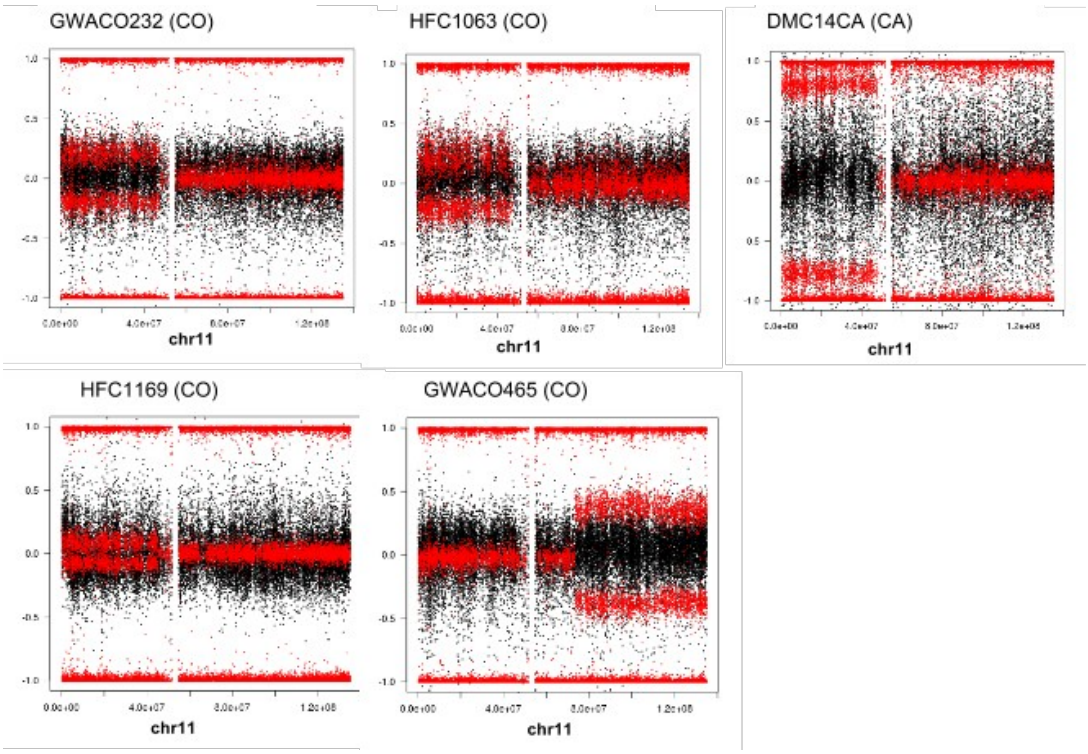


Fig 2. Rare mosaic events of 11p in three controls out of four. Additional control affected with 11q mosaic event.

No mosaic event was more prevalent in cases than controls.

CNV

After filter-out the outlier samples, we end up with 1735 samples. 770 of them cases and 965 controls (individuals with mosaic events and contaminated filtered). We got 7,400 events in pennCNV (filtering also by >= 25 events per sample)

Finally the distribution of CNV was the following :

Large CNVs (> 100 probes) 1.7 % (128 out of 7400, manually curated and 2 discarded).

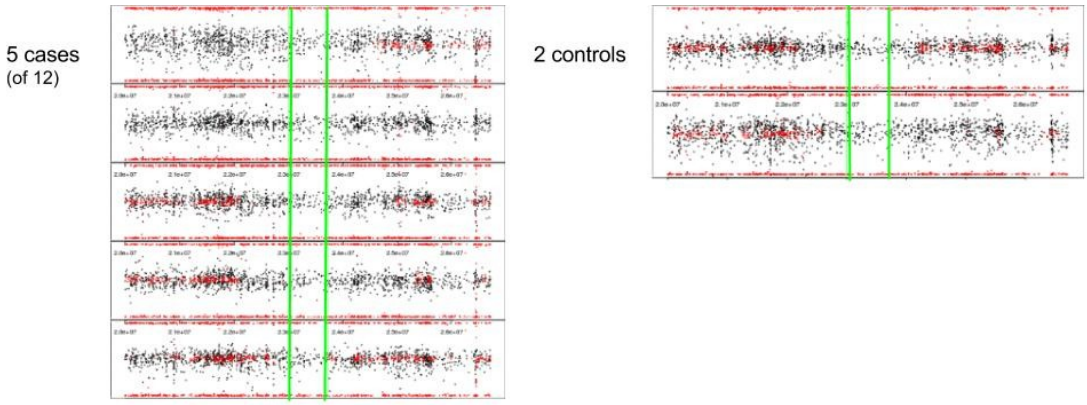
Small CNVs (< 100 probes) 7272 calling events (without apply any astringent filter).

No valid CNV surpassed bonferroni test. Other positive callings were discarded, as they are likely to be artifacts. Usually the majority of them are located at the end of the chromosomes or in the centromers or in zones rich on segmental duplication regions.

IBDs

Performing the case control study of IBD regions, although there is no significant result (applying Bonferroni), we got one significant result.

Region coordinates	Ca	Co	OR	95%CI	p-value
chr3:23639706-23659405	12/806	2/1026	7.77	2.08-54.5	0.0009



This zone was surrounding UBE2E2 and is related to autoimmune diseases (*arthritis and systemic lupus*) performing the role of regulating the inflammation process via ubiquitination. UBE2E2 (ubiquitin-conjugating enzyme E2E2) is a protein-coding gene.

Exploring those available SNPs in the region, three of them appear to be significant according to Bonferroni with a threshold of 0.05/47= 0.001 (the p-values are not corrected).

estimate	lower	upper	p-value_chi	Snp.name	Allele value	Num CA	Num CO	Total CA	Total CO
1.78203	1.33037	2.39378	0.00009	rs4619807	GG	113	89	872	1155
1.45954	1.18597	1.79652	0.00034	rs6550734	CC	235	233	872	1155
1.36916	1.14395	1.63899	0.00060	rs6771206	GG	382	419	872	1155

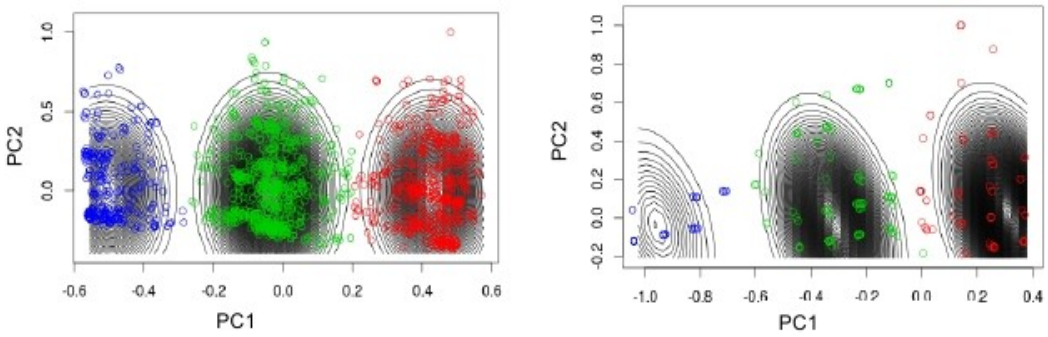
Snp.name	Allele value	Num (IBD)	CA% num (IBD)	CAPopulation val allele freq %
rs4619807	GG	7	0.583	7.7
rs6550734	CC	9	0.75	20.17
rs6771206	GG	9	0.75	36.27

Table 2. Significant available SNPs in IBD region.

Inversion association study

Performing an association, correcting by multidimensional scaling of the population and sex, 3 putative inversion shows a statistically significant correlation over the selected 1832 samples (806 cases versus 1026 controls):

	OR	pvalue	freqI	model	coordinates
7q35	0.78	0.0215	0.45	dominant	chr7: 142988068-143509153
16p11	1.26	0.0031	0.24	additive	chr16: 28256775-28562004



In Fig 4 the putative inversion 7q35 forms a clear three clusters, suggesting a clear two main haplotypes of the inversion, and the non-inverted allele is the risk allele.Whereas the 16p11 is already known and characterized [8], and it shows a clear correlation against the inverted-allele.

Discussion

We used a database of 2561 European trios in order to define the accuracy of the haplotype prediction by *InvClust* for these inversions. In some inversions the variability between population could be considerable, but the patterns of both inversions are quite similars, despite to have differences on their frequencies. We detected a Mendelian error rate of 12/2561, and 10/2561 for the 7q35 and 16p11 respectively. Those results that makes us confident about the accuracy of predicted genotypes.

Using Hapmap and Estonian population datasets, we established a correlation of the predicted inversion alleles with gene expression of regional genes in blood. The effects

of 16p11 over gene expression were already described in literature [8]. 16p11 inversion shows a significant protective effect on the risk of asthma, that become stronger with the co-occurrence of asthma and obesity. The non-inverted inversion allele regulates the gene expression positively with *TUFM*, *SNPS*, *SULT1A1* and negatively with *SULT1A4*, *CCDC101*, and heterozygous have positive effect over *IL27*.

On the other hand, 7p35 inverted allele correlates **negatively** with *OR2A9P*. *OR2A9P* is an olfactory receptor protein member of a large family of G-protein-coupled receptor (GPCR).

Bibliography

[1]. Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *PloS One*, *7*(7), e40224. doi:10.1371/journal.pone.0040224

[2]. Salm, M. P. a, Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., … Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research*, *22*(6), 1144–53. doi:10.1101/gr.126037.111

[3]. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–75. doi:10.1086/519795

[4]. González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., … Pérez-Jurado, L. a. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *American Journal of Human Genetics*, *94*(3), 361–72. doi:10.1016/j.ajhg.2014.01.015

[5]. Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. *BMC Bioinformatics*, *13*(1), 28. doi:10.1186/1471-2105-13-28

[6]. HapMap project [http://hapmap.ncbi.nlm.nih.gov/]

[7]. González, J. R., Rodríguez-Santiago, B., Cáceres, A., Pique-Regi, R., Rothman, N., Chanock, S. J., … Pérez-Jurado, L. a. (2011). A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics*, *12*(1), 166. doi:10.1186/1471-2105-12-166

[8]. González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., … Pérez-Jurado, L. a. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. American Journal of Human Genetics, 94(3), 361–72. doi:10.1016/j.ajhg.2014.01.015

6. DISCUSSION

The main focus of this thesis work was to study chromosomal mutations that have not been yet carefully analyzed, such as submicroscopic inversions or other structural variants present in mosaicism, and test their putative involvement in the pathogenesis of some model complex diseases. The main reason of the under-exploration of these genomic variants has been the difficulty to detect and genotype them in large cohorts. Our aim was to use optimized methods for high throughput detection and genotyping developed by our group and collaborators to be able to reanalyze and exploit already available SNP genotype (and phenotype) data of several multifactorial disorders, especially of neurobehavioral [Autism Spectrum Disorder (ASD) and Schizophrenia Spectrum Disorder (SSD)] and autoimmune [Rheumatoid Arthritis (RA) and Ulcerative Colitis (UC)] diseases. The rational is based on the main hypothesis that studying common genetic variants in large cohorts, it would be possible to find biomarkers that explain part of the high genetic component of those disorders that has been demonstrated by twin and other epidemiological studies.

During the last 10 years, multiple GWAS case-control studies have been done with common SNPs, motivated with the idea that common disorders should be caused by common genetic variants. In other words, variants that are common in population should be significantly overrepresented in cases compared with healthy controls. However, GWAS with SNPs have not met expectations in most cases; indeed most of them failed on replication in different cohorts. Only in few complex disorders, such as Schizophrenia (Bergen, S. E., & Petryshen, T. L. (2012)), GWAS have shown better replication results. Other approaches focused in searching for highly penetrant but rare genetic variants, such as deleterious point mutations, small insertions and deletions, as well as large Copy Number Variations (CNVs) (Malhotra D. & Jonathan Sebat J. (2012), Bergen S & Petryshen T. (2012)) have yielded significant results, but globally explain a small proportion of the genetic component of those disorders.

In ASD for example, all discovered changes until today, genetic variants including recurrent CNVs, rare-variants, *de novo* CNVs, loss of function mutations and other, explain the cause of roughly ~20-30% of the cases. Therefore, despite all efforts

and genomic approaches available, the cause of the majority of affected ASD cases (~70-80%) remains unknown.

We could depict a quite similar picture for most of multifactorial disorders. Hundreds of potential genetic changes are found in most of the most common complex diseases, but they individually account for a relatively small amount of the heritability of those disorders, even when additive small effects of multiple variants are taken into account.

Where is the missing heritability?

Several studies discuss this topic (Maher, B. (2008), Manolio, T.a. et al (2009), Zuk, O et al. (2012)). If our main starting hypothesis is that we could find a portion of measurable heritable trait examining the genetic code, the main limitation is the used technology. For example, using SNP arrays, it is only possible to detect those variants present in the SNP library, as well as CNVs in regions covered by the probes. On the other hand, SNP arrays only check a subset of common variants and do not capture rare variants other than large CNVs. Next generation sequencing may capture more variation but is also limited to the targeted sequence. Using whole-genome sequencing (WGS) technologies based on relatively short reads and an incomplete assembled genome, the detection rate improves. However, there are still limitations as WGS is still difficult to apply to large cohort and may not detect some variations such as chromosomal inversions and changes in complex regions such as segmental duplications. The best solution could have larger reads, but it is not possible with the available technologies. Other source of variability could be hidden by epigenomic marks that could induce to the silencing of certain genes, and part of the variation may require the study of gene expression to detect variant isoforms and splicing events.

Using the available technology, a clever mining of the data using informatic tools improved the ability to identify some of the variations of copy-number (duplications, deletions, CNVs and mosaic events) that were initially undetected. Therefore it makes a lot of sense to try to perform integrated analysis, using all available data such whole-sequencing genome, epigenomic marks and transcript analysis, and in any case it is a guarantee to found all variation. But of course, not always is possible have all those resources to explode.

Given that a remarkable amount of information has already been obtained with SNP arrays in multiple GWAS studies and can be available through public resources such as dbGap, it is important to extract as much information as possible of those resources. So we expect to find some sets of potential biomarkers that alone have not enough power to pass statistical test, but altogether be significant. Therefore, we asked the question whether inversions are one of those common variants of lower penetrance that add a relevant risk factor to take into account. Could inversions help us to understand little more the mechanisms involved in certain complex disorders?

This idea has been previously explored in a few published works, from our research group and others, with promising results (Salm et al (2012), De Jong, S. et al (2012), Cáceres, A. (2015), Ma, J. et al (2014)). An illustrative examples the 8p23.1 inversion was associated with systemic lupus erythematosus (Salm et al (2012), and a novel 16p11.2 inversion was found to provide a join susceptibility to asthma and obesity (González, J.R.(2014)). Other results included uncharacterized inversions such as Ma J et al (2014) that found 5 putative inversions that could be related to psoriasis.

Strong and weak points in the inversion detection used methods

The central core of this PhD was searching variability hidden in common submicroscopic inversions (~0.5-4.5 Mb), and other variants, that could be detected and genotyped from SNP microarray data. Actually there is no gold standard method to characterize and genotype submicroscopic inversions. As commented in the method section, the core strategy that in a first step we scan for region susceptible to contain an inversion (localize breakpoints), and then check if there is an inversion pattern (at least two main haplotypes). We mainly used *inveRsion* and *invClust* algorithms (see Methods). *InveRsion* algorithm computes LD between two contiguous blocks of SNPs that flanks a potential breakpoints and tests if there is a disruption in the expected LD between the two blocks that could reveal an inversion. An alternative tool could be *invertFREGENE,* but it seems not to perform as well as *inveRsion* with inversions with high frequencies. The reason is that inversions with higher frequencies tend to be older, which may then present

higher within-population variability than can impact the prediction. On the other hand, *inveRsion* method performance is impacted mainly by the length and frequency of the inversion as well the SNP density. In general, more SNP density gives better accuracy.

As commented, *inveRsion* method gives the possibility of scanning genome (SNP array) with predefined window sizes in order to search the putative inversion breakpoints. This first step is quite robust but resource costly, and it gives false positives. As a result we get a list of putative inversion breakpoints regions.

Once the putative inversion regions are found, the next step was to check if there are patterns of inversions. For example, it is possible to perform just a *Multi-Dimensional Scaling (MDS),* and if three clusters are found (3 genotypes), it is likely to be an inversion (the presence of two clear haplotypes). The main argument is that we are detecting two clear haplotypes in a region that *inveRsion* method previously detected putative breakpoints of inversion (changes in LD between the contiguous SNP blocks). The added value in using *invClust* tool is that it not only can infer the genotype for available data, but it also can estimates a confidence level in the classification of individuals. *InvClust* is not a perfect method, and therefore it is not enough guarantees that there is a submicroscopic inversion there, but we gain in confidence if the breakpoints are delimited with *inveRsion*. In fact, with this method, our group predicted a total of 174 putative ancient submicroscopic inversions in the human genome, including the few already known. Of them, 66 inversions showed a remarkable correlation with differentially expressed genes in blood samples. In posterior studies we also checked the degree of concordance of Mendelian errors in different trio datasets (with more than 1,500 trios), and we assessed that at least there are 24 putative inversions with high degree of concordance. An advantage of the use of trios is that one can check for coherence with Mendelian laws, providing more confidence to the genotype prediction.

Reasons for poor clustering are multiple, including the arbitrary disposition along the PC1 and PC2 components. Although *invClust* is more robust than other methods, there are also limitations due to mixed populations and population stratification as a confounding factor. In any case, all these potential sources of errors could be in some cases identified checking the available trios, but there is no guarantee of errors-free.

If in a given region there was positive signals from *inveRsion* and *invClust* with 3 clear clusters, the consistency between different methods provide more confidence of an underlying biological basis, likely an inversion. There are unusual situations in which more than 3 clusters are detected even with the previous validation of an inversion, such as for the 16p11.2 or 15q24.2 regions. In those cases, different methods were used to infer the clusters (f.eg. *kmeans* algorithm or methods implemented in R packages such as *mclust*, *htclust*), but it would be necessary to validate the predicted genotypes to demonstrate that selected haplotype patterns have indeed been shaped by inversions, likely more than one at the same locus. The main problem is that it is not a trivial task to infer it manually, and for example using FISH techniques, sometimes it is a difficult task for particularity of the region. The flanking large segmental duplication regions, full of repeats motifs, makes difficult use molecular techniques in some cases (e.g. limitations in the length of primers). Ideally, a whole-genome sequencing with 4,000,000 length reads were am optimal method, but the technology is not enough mature (usually in next-generation sequencing technologies reads have less than 900 bp) [3].

And finally, tagging a certain inversion with few SNPs it is not always possible, and you need hundreds of SNPs (e.g. the 8p23 inversion). When an inversion is correctly tagged, it is easier to perform studies if those SNPs are available.

Every inversion has its own nuances and it is not a trivial task prof that certain region is in fact an inversion. Therefore it is difficult perform this kind studies.

Association of 17q21.31 and 8p23 with neurocognitive complex disorders

In this PhD, we analyzed submicroscopic inversions with three available databases with ASD affected samples with their parents. All together conforms 5,089 families, the majority of them with complete trios (European and non-European ancestry), and with 1,363 healthy brothers. Out of the four genotyped inversions, we found very few errors in Mendelian inheritance reinforcing the accuracy of the haplotype predictions. When testing the transmission disequilibrium we found that the H2 inverted allele (I) of inv17q21.31 was found significantly over-transmitted to ASD probands mainly in multiplex families, as well as in ASD probands who did not fulfill criteria for strict

[3] https://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods

autism, being verbal and with intellectual quotient (IQ) above 80, from either simplex or multiples families. The association of ASD with over-transmission of the H2 haplotype was not replicated in SSC, but UMSGARD an over-transmission of the H2 allele was also noticed, although it did not reach significance. A meta-analysis revealed significant evidence of an overall 6.02% over-transmission of the H2 allele by 5.58% mostly of paternal-specific (9.91%, p=5.3e-04).

Despite this clear overall association of ASD with paternal H2 over-transmission, there were remarkable differences between studies with completely null results in SSC. One possible explanation for the discrepancy can be related to the different composition and phenotype of the samples in the three studies analyzed. While AGP and also MSGARD are enriched in multiplex families and include a significant proportion of ASD cases that do not fulfill criteria for strict autism, SSC is restricted to cases of simplex families with the great majority fulfilling criteria for strict autism. In fact, when using only AGP trios in which the case had a strict autism diagnosis, the TDT was not significant. Therefore, the H2 or I allele at inv17q21.23 appears to be associated with increased risk for a specific subphenotype of ASD with preserved verbal communication and high IQ, and is more prevalent in families with more than one case of ASD.

The I-allele of the 4.5Mb inv8p23.1 was also over-transmitted in the three independent ASD datasets (7.1%, p=2.5e-05). The contribution of maternal and paternal transmissions was similar and in the same direction, with no gender bias. A significant increment of the association was observed in verbal ASD with low IQ (16% over-transmission, p=0.0059). A case-control study validated the findings of significantly increased risk of ASD associated with the H2 allele of inv17q21.31 (OR=1.14) and the I-allele of inv8p23.1 (OR=1.22). Interestingly, H2-haplotype at inv17q21.31was nominally associated with protection for SSD (OR=0.86).

The inv17q21.31 has a length about ~970 Mb located at 17q21.31, and it is composed with two main alleles the reference allele (H1) and the inverted (H2). The inverted haplotype appeared about 3 million years old, and there are few evidences that the inverted haplotype had recombined forming a complete independent haplotype compared with the non-inverted allele. Their frequency varies according population from Europeans (~22%), but few common in Africans (6%) and Asiatic population (1%). The origin of the formation of common ancient submicroscopic inversion events in most of the cases is due to the presence of two flanking segmental duplications, and certainly inv17q21.31 is no exception. In fact, in segmental duplications there are

several protein coding genes such as *ARL17P*, *PLEKHM1* and *LOC474170*, that consequently they could vary in copy number. A surprising feature of this inversion, is that is under selection in Europeans. An analysis performed in Island population shows that the women carrying the inverted-haplotype have in mean more child, than those women with the non-inverted. Eight different sub-haplotypes were cataloged, five haplotypes for the non-inverted allele (H1), and three for the inverted allele (H2) (Steinberg et al (2012)). The main differences among them, was the differences of copy-number of *KANSL1* gene. It was possible to differentiate all H2 sub-haplotypes, but not the remaining H1 sub-haplotypes with the available SNPs. Nevertheless, no significant proportion was observed in the frequency distribution of each haplotype with european samples.

This is a complex region in genome in which have been found several recurrent deleterious rearrangements, such as the 17q21.31 microdeletion syndrome, characterized by intellectual disability, hypotonia and distinctive facial features caused by an haploinsufficency of *KANSL1* (Koolen, D. a et al (2012), Zollino, M. et al (2012)). The other is the 17q21.31 microduplication that causes behavioral problems and poor social interaction.

Other studies suggest that inversion 17q21.31 *MAPT* H1 haplotype (non-inverted 17q21.31 allele) could be associated with progressive supranuclear palsy, corticobasal degeneration and Parkinson and Alzheimer's disease (Bakker et al (1999), Skipper et al. (2004), Myers AJ (2005), Webb A et al (2008), De Jong, S. et al (2012)).

The inverted allele of 17q21.31 associated overexpression of *CRHR1* in blood and brain tissues, especially in cerebellum (p-value 1.06e-3). Corticotrophin releasing hormone receptor 1 is a major regulator of the hypothalamic-pituitary-adrenal axis and mediates stress response. This suggest that more receptors of *CRHR1* are produced in the carriers of this allele, and might confer susceptibility to ASD through facilitating the deleterious effect of prenatal and early infancy stressful events on neuronal synapses.

MAPT shows an overexpression with the non-inverted allele of 17q21.31 in cerebellum (1.53e-15). It is not clear if it is related or not with ASD. *MAPT* has been reported as a risk factor for supranuclear palsy, corticobasal degeneration, neurodegenerative tauopathy, Alzheimer, and Parkinson diseases (Webb et al. 2008, Li et al. 2014, Mayers

et al. 2005, Zabetian et al. 2007). The H1 haplotype is referred as the non-inverted allele, and the H2 to inverted allele. GWAS revealed variants in the region associated with cranial volume in infants of the general population (Taal et al. 2012). Moreover, *CHRN1* SNPs in the H2 haplotype respond better to inhaled corticosteroid in asthma (Tantisira et al. 2008). H2 is also a protective factor through gene-environmental interaction for sexual abuse-associated alcohol dependence (Nelson et al. 2010).

KANSL1 is under-expressed in temporal cortex (p-value 7.83e-3), and could be an ASD risk gene. *KANSL1* haploinsufficiency has been reported as the main or single cause of the 17q21.31 deletion syndrome, since the novo loss-of-function mutations in *KANSL1* (also called *KIAA1267*) cause a full del(17q21.31) phenotype (Zollino, M. et al (2012)). This disorder is characterized by highly distinctive facial features, moderate-to-severe intellectual disability, hypotonia and friendly behavior.

ARHGAP27 is underexpressed in in frontal cortex as well as cerebellum (similar effect p-values ~3.5e-2). This gene was reported as tumor suppressor gene, as their genetic alterations lead to carcinogenesis through the deregulation of Rho/Rac/ Cdc42-like GTPases. But we cannot discard that have some kind of role in risk factor for ASD.

The inversion 8p23.1 is a well-characterized inversion that also shows a similar structure to 17q21.31 with two large blocks of segmental duplications flanking a 3.4 Mb single copy interval in the short arm of chromosome 8. Its frequency varies among populations, in Europeans is more frequent the inverted allele with ~56 %, with an increased value in Africans ~80%, while non-he inverted allele is the predominant in the Asian population, and almost the only one in the Oceanic population. SNPs at genes inside this inversion region have been shown associated to susceptibility to Lupus, Asthma as well as psychiatric disorders. Among the reorganizations described in literature affecting 8p23.1 we found tandem duplications, inverted duplications, deletions, pericentric and paracentric inversions and translocations. It is a region characterized by their instability. In fact, the complexity of segmental duplications of that region makes it difficult to accurately define the breakpoints.

The 8p23.1 region harbors structural polymorphisms mostly at the flanking segmental duplications, and many of them are CNVs. The defensin gene clusters located in the segmental duplications have special interest due to its probable relation to innate immunity and cancer. This copy-number variation is an essential and some alleles could be related to some diseases, especially if the genetic dosage influences genic expression (Mars WM et al (1995), Zhang L et al (2002), Ganz T et

al (2003)). The left segmental duplication zones, is characterized by a cluster of defensin genes such as *DEFB104A*, as well as sperm maturation genes such as *SPAG11B*. Located in the right segmental duplication region, near *CTSB* gene, it is found another cluster of defensin genes such as *DEF1A3* and pseudogene *DEFTP*.

The inversion allele correlates significantly with changes in gene expression of the following genes in blood and brain tissues: *BLK* (C8orf13), *CTSB*, *FDFT1* and *SOX7*.

The *BLK B-lymphocyte kinase* of the tyrosine kinase family, is a protein-coding gene related to autoimmune disease susceptibility. Non-receptor tyrosine kinase involved in B-lymphocyte development, differentiation and signaling. B-cell receptor (BCR) signaling requires a tight regulation of several protein tyrosine kinases and phosphatases, and associated co-receptors (RefSeq). *BLK* is a class I MHC mediated antigen processing and presentation and, as well as a susceptibility gene of some autoimmune diseases such as Systemic lupus erythematosus (SLE) and Rheumatoid Arthritis (RA). Additionally, The protein also stimulates insulin synthesis and secretion in response to glucose and enhances the expression of several pancreatic beta-cell transcription factors, and in fact it is related to diabetes of type 11. And the non-inverted allele of 8p23.1 produced a reduced expression in blood producing a reduction in *BLK* gene expressing, provoking a susceptibility risk factor for SLE and RA.

Our results shows an increase of their expression in brain tissue, special in pons (p-value 5.28e-5) with the inverted allele, suggesting a possible relation with ASD. The risk factor should be with the decrease of *BLK* expression (Simpfendorfer, K. R. (2012)). Therefore it is likely that the non-inverted allele should be related with autoimmune disorders.

CTSB cathepsin B is a precursor (proteolysis protein) of beta amyloid protein. Amyloid fibrillar form is primary component of amyloid plaques found in the brains of Alzheimer's disease patients, therefore it is a candidate gene of neurodegeneration. Our studies point out that an under-expression of this gene with the inverted allele (p-value 2.25e-2). Their function seems to be regulator of synapse formation, neural plasticity and iron export, and therefore we can point *CTSB* as a possible risk factor for ASD. Additionally it was found that was a novel candidate gene for emotionality in mice (Czibere et al 2011), that could support more this idea.

MSRA it is a highly conserved protein that carries out the enzymatic reduction of methionine sulfoxide to methionine. Human and animal studies have shown the highest

levels of expression in kidney and nervous tissue. The protein functions in the repair of oxidatively damaged proteins to restore biological activity. The inverted allele of 8p23 shows an under-expression of MSRA, especially in pons tissue (p-value 9.91e-5), and curiously increase its expression in cerebellum (6.19e-3).

FDFT1 gene encodes a membrane-associated enzyme located at a branch point in the mevalonate pathway. The encoded protein is the first specific enzyme in cholesterol biosynthesis, catalyzing the dimerization of two molecules of farnesyl diphosphate in a two-step reaction to form squalene (RefSeq). *FDFT1* was previously associated in fibrosis progression in patients with chronic hepatitis C. (Stättermayer AF et al (2014)). Inverted allele 8p23 produces an increase of *FDFT1* expression in cerebellum and frontal cortex tissues (p-value ~2.7).

Association of 15q24.2 inversion with IQ trait

We found positive association with a verbal and non-verbal intelligence quotient (IQ) in 2,735 children of European ancestry. But in this case, although there is an already detected 15q24.2 inversion, we found that it was likely to have three haplotypes instead of two. And posterior results show less Mendelian errors assuming three haplotypes than two. Additionally, the *inveRsion* we got an additional breakpoint that could support the fact to have an inversion inside of other. We confirmed this model with a larger dataset trios using AGP dataset that supports our genotype inversion model (only 6 Mendelian errors out of 2562).

The chromosome bands 15q24.1-15q24.3 contain a complex region with numerous structural variations that include microduplications and microdeletions, both of which 4 have been linked to intellectual disability, speech delay and autistic features.

IQ is a complex trait with a heritability in adult twins of 70-80% and child twins of 45% (Plomin, R. et al. (1994),Deary Ian J., Batty G. David (2007)). Most IQ tests are constructed so that there are no overall score differences between females and males. Popular IQ batteries such as the WAIS and the WISC-R are also constructed in order to eliminate sex differences. The 15q24.2 inversion study was replicated on three different databases, joining altogether into a single meta-analysis. Although different IQ test were used in those different databases, all measure almost the same dimensions of IQ and using very similar tests based on the standard WAIS and WISC-R such as Wechsler

Intelligence Scale for Children III (WISCIII) and Dutch battery [TaaltestvoorKinderen (TvK)] and others, taking into account the proper correction by age.

Homozygosity for the Ia and Ib haplotypes were both associated with non-verbal IQ: 1.5 point gain for Ia (p-value=0.0004) and 2 point loss (p-value=0.001) for Ib. In addition, the NI allele strongly correlated with higher expression levels of *MAN2C1* in blood (p-value<10-15) and brain (p-value=0.02) while the Ib allele significantly down-regulated *SNUPN* in both tissues (blood p-value=0.01, brain p-value=0.01). Our data indicate that common polymorphic inversion-related haplotypes at 15q24.2 with variable population distribution influence human intelligence most likely by regulating local gene expression.

MAN2C1 and SNUPN were consistently up-regulated at the NI allele and down-regulated at the Ib configuration, both in blood and brain tissues. MAN2C1 has been shown to have a dual function. MAN2C1 encodes the alpha-mannosidase, class 2C, member 1 that has been shown to regulate protein N-glycosylation and apoptosis. The N-Glycoproteome maps mainly to blood but the highest amount of organ specific N-glycosylation sites in mice has been observed in the brain (Zielinska DF et al (2010)). As the MAN2C1 gene is highly expressed in hippocampal formation, its differential regulation by the different haplotypes might be related to the observed differences in cognitive function. Over-expression of MAN2C1 leads to protein underglycolysation and up-regulation of the degradation of unfolded glycoproteins (Bernon C et al (2011)). The attachment of glycans to some proteins is important for their correct folding and/or stability. N-glycans cover diverse biological functions in the nervous system, ranging from the essential to the modulation of development and neural transmission, which in turn can affect plasticity and memory formation (Scott H & Panin VM (2014)).

Association of unusual chromosomic variants with UC

Ulcerative Colitis (UC) is subcategory of inflammatory bowel disease that causes inflammation and ulcers in the colon. UC have an incidence of 1.2-20.3 cases for every 100,000 persons per year, with a prevalence of 7.6-246 cases per 100,000 per year (Colitis, U. (2003), Danese, S., & Fiocchi, C. (2015)). It changes their prevalence frequencies based on different populations, e.g. in European population

is rare (less than 1%), whereas in Japanese population it is highly prevalent (20%-25%). Most of the heritability of UC (>70%) is still unknown.

No clear association of UC was found with mosaic events, although a rare 11p mosaic event appear in three out of four controls and only one case. Curiously the percentage of cells affected by mosaicism was 70% in the cases, whereas all the remaining controls have less than 40% (calculated based on Bdev). We know that mosaic events are supposed to be rare events. There is a very rare event found mosaic events on children, but in advanced ages are becoming more frequent. In any case, our collaborators were unable to give more information regarding the age of those samples, and more information if, during larger period of time, developed UC.

The analysis of CNVs do not give positive results, and CNV was more prevalent in cases versus controls.

The analysis of overlapping regions of homozygosity returned one significant region on 3p23.3 (OR 7.77 CI 95% [2.08-54.56]). Twelve cases show an overlapping in this homozygosity region against two controls. Exploring the available SNPs in this region the only putative recessive allele was rs4619807 (OR 1.78 p-value 0.00009). It was not reported previously for UC, maybe because it does not surpassed Bonferroni GWAS correction (it was on the limit). Nevertheless is highly suggestive that this region is a potential source of missing heritability for UC. This region contains the UBE2E2 and UBE2E1 genes coding for the ubiquitin-conjugating enzymes E2E2 & E2E1. UBE2E2 has been previously related to some autoimmune diseases (arthritis and systemic lupus), but not for UC.

Two inversions appear to be correlated with UC, the 7q35 (OR 1.28 non-inverted allele p-value 0.021) and 16p11 (OR 1.26 inverted-allele p-value 0.0031). The 7q35 is a putative inversion, and already not fully characterized, although it shows low Mendelian errors. On the other hand 16p11 was an already studied submicroscopic inversion. The previous results show that non-inverted allele of 16p11 inversion shows a join susceptibility for asthma and obesity (all together). Curiously, it is the opposite fact in UC, so the inverted allele is the risk factor. On the other hand, 16p11 shows no correlation with ASD nor SSD available samples in linkage and association studies. Regarding autoimmune available complex disorders, 16p11 inverted allele shows an association with Ulcerative Colitis (OR 1.26 inverted-allele p-value 0.0031).

Expression effects were observed on single-copy genes located within the inversion (*IL27* and *CCDC101*) and immediately proximal interval (*TUFM* and *SPNS1*), as well as on multiple-copy genes located in the flanking segmental duplications (EIF3C, *SULT1A1*, and *SULT1A4*). In the case of *TUFM*, *SPNS1*, *EIF3C*, *SULT1A1*, and *APOB48R*, expression levels were higher in I alleles (González, J. R. et al (2014)). And all those genes could be potential susceptibility risk factors of UC.

Some of biological explanation of some of this differential expressed genes such as *IL27* and *TUFM* could be the following.

IL27 is one of the critical cytokines that mediates between the innate and adaptive immune system. *IL-27* has been implicated in the pathogenesis of many autoimmune inflammatory disorders, including rheumatoid arthritis, psoriasis, multiple sclerosis as well as Crohn's disease, and ulcerative colitis. *IL27* regulates T and NK-T cells activity and also mediates inflammation.

TUFM has also already been related with inflammatory cytokine cell-signaling pathway. *TUFM* promotes autophagy when associates to ATG5-ATG12 complexes and with ATG16L1 (Vojo Deretic et al, Autophagy in infection, inflammation and immunity, 2013).

Association of unusual chromosomic variants with RA

Rheumatoid arthritis (RA) is an autoimmune disease characterized by persistent synovitis, systemic inflammation and autoantibodies. Genetic factors account for 50 % of the risk to develop RA, and heritability was estimated to be about ~60%. The largest contribution was major histocompatibility complex with a recurrent risk for siblings ~1.8, followed by few minor variants. Nevertheless, most proportion of heritability is still unknown.

Compared with UC, Rheumatoid Arthritis (RA) is tough to diagnose. In fact, the main problem is to differentiate RA from other inflammatory arthritis, such as psoriatic arthritis, lupus, reactive arthritis, spondylopathies, among others. Although there is no singular test for diagnosing RA, using several of them, could help to elucidate confidently which type of arthritis it is. Among those tests, besides a full physical exam, there are imaging tests (X-ray or MRI), blood tests to check inflammation, joint fluid tests, and arthroscopy.

Therefore the confidence to assess UC and RA and classify homogeneous patients is high. Nevertheless, neurocognitive disorders such as ASD or SSD is not so relatively easy, and it is thought that exist more heterogeneity.

No mosaic events were statistically found in RA. And despite some CNVs surpassed Bonferroni correction, were discarded to be possible artifacts.

The case-control association test revealed a single region located on chr4q32.1 that was found in homozygosity region significantly more enriched in cases than in controls. The minimal interval of overlap (chr4:160244776-161281594). This region was overlapped by 18 cases (of 990) and 2 controls (of 1078) with an OR 9.32 [CI 95% 2.66-63.46].

The biological relevance of this region with RA, could be that contain RAPGEF2 gene, that previously was associated with arthritis and autoimmune disorders. It is a G-Protein RAS works as signal transduction in the cell working as a molecular switch that activates a kinases cascade. Searching in this region, the most suggestive recessive SNP was rs1158264 although it does not so strong statistical power (chi square p-value 0.049) and relatively weak OR 1.2 [1.00-1.44]. Logical in any case, because it was not previously been described.

Future directions

The main source of data for this thesis was the use SNP arrays. The advantage of this mature technology is that allows to detect common and rare variation (CNVs, mosaic events) with a relatively low cost. And a precisely a lot of studies and public datasets are available.

Of course seems inevitable in the incoming years, that DNA arrays were displaced by NGS (Thomas LaFramboise (2009), Shendure (2008)). The cost of next-generation sequencing (NGS) technologies is dramatically decreasing in last years, but SNP arrays seem a reliable option in many situations. The already studied biomarkers and the fact that the technologies include high-density SNP array panels, makes this technology the most used in clinics (diagnose of leukemia, prenatal and postnatal detection panels of mental disorders, etc.).

A main disadvantage is precisely that old microarrays have few resolution (low SNP density), and sometimes we were not able to genotype inversions because of that.

Indeed, with more resolution of SNPs is more likely to correctly characterize an inversion, as well as tag it with few SNPs. Therefore, clear future direction is to get in any case, high-density microarray panels. And if it is possible, whole-genome sequenced samples. In fact the ideal case should be the use of a technology that allows reads of ~5Mbp, but this technology is very immature (such as Pacbio). The usual read lengths are larger than 900 bp.

With these techniques, in the incoming years it is likely to discover new inversion variants. Therefore the creation of a database to store all this information should be necessary. Not only the inversions, so their effects at gene expression level, susceptible disorders, etc. Actually there is an initiative, and already exist a database of inversions, InvFEST (Martinez-Fundichely et al., (2014)) that integrates information of polymorphic inversions in human genome. But there is still a huge work to do in order to make all this information more accessible (using APIs, etc.), and let us to explore this amount of hidden possibilities.

7. CONCLUSION

In this thesis we found evidences that ancient submicroscopic inversions could explain a portion of missing heritability of immunological and neurocognitive complex disorders.

Validating our starting hypothesis we arrived to the following conclusions:

1) We improved a protocol and tools to detect and improve the characterization of submicroscopic ancient inversions (0.5 – 5Mb) using SNP array technologies. That is, using the *inveRsion* and *invClust* algorithms to identify putative inversions, and perform additionally on trios available datasets as an attempt to improve characterization of some inversions.

2) With the previous protocol we identified 174 putative inversions computationally using SNP arrays. From those, 66 inversions have correlation with gene expression in blood. And from those 66 inversions, we identified 24 of them shows low Mendelian errors and we have strong hypothesis that might be inversions.

3) Performing linkage and association studies, we found that 17q21.31 and 8p23 inversions have association with neurocognitive disorders such as ASD and SSD. Concretely inverted alleles of 17q21.31 and 8p23 with ASD, non-inverted allele with SSD, while it was already found that non-inverted allele of 8p23 was associated with Systemic Lupus Erythematosus (SLE) (Salm, M. P. et al (2012)). The inverted allele of 16p11 inversion was also found a joint susceptibility factor by asthma and obesity but as risk factor for UC.

4) The North Indian population study of RA and UC give positive results analyzing large regions of homozygosity.

Therefore we add our sand of grain to the proportion of some complex disorders, and we are able to found in all of them relevant biological hypothesis.

8. BIBLIOGRAPHY

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M. S., Tomasini, L., … Vaccarino, F. M. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature, 492, 438–442. doi:10.1038/nature11629
Agerbo, E., Sullivan, P. F., Vilhjálmsson, B. J., Pedersen, C. B., Mors, O., Børglum, A.

Agerbo, E., Sullivan PF, Vilhjálmsson BJ, Pedersen CB., … Mortensen, P. B. (2015). Polygenic Risk Score, Parental Socioeconomic Status, Family History of Psychiatric Disorders, and the Risk for Schizophrenia: A Danish Population-Based Study and Meta-analysis. JAMA Psychiatry, 1–7. doi:10.1001/jamapsychiatry.2015.0346

Aguado, C., Gayà-Vidal, M., Villatoro, S., Oliva, M., Izquierdo, D., Giner-Delgado, C., … Cáceres, M. (2014). Validation and Genotyping of Multiple Human Polymorphic Inversions Mediated by Inverted Repeats Reveals a High Degree of Recurrence. PLoS Genetics, 10(3), 14–22. doi:10.1371/journal.pgen.1004208

Alam, S., & Kelleher, S. L. (2012). Cellular mechanisms of zinc dysregulation: A perspective on zinc homeostasis as an etiological factor in the development and progression of breast cancer. Nutrients. doi:10.3390/nu4080875

Ammitzbøll, C. G., Kjær, T. R., Steffensen, R., Stengaard-Pedersen, K., Nielsen, H. J., Thiel, S., … Jensenius, J. C. (2012). Non-synonymous polymorphisms in the FCN1 gene determine ligand-binding ability and serum levels of M-ficolin. PloS One, 7(11), e50585. doi:10.1371/journal.pone.0050585

Argente, J., Flores, R., Gutiérrez-Arumí, A., Verma, B., Martos-Moreno, G. Á., Cuscó, I., … Pérez-Jurado, L. A. (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. EMBO Molecular Medicine, 6, 299–306. doi:10.1002/emmm.201303573

Arimura, Y., Isshiki, H., Onodera, K., Nagaishi, K., Yamashita, K., Sonoda, T., … Shinomura, Y. (2014). Characteristics of Japanese inflammatory bowel disease susceptibility loci. Journal of Gastroenterology, 49(8), 1217–30. doi:10.1007/s00535-013-0866-2

Ashwood, P., Wills, S., & Water, J. Van De. (2006). The immune response in autism: a new frontier for autism research, 80(July), 1–15. doi:10.1189/jlb.1205707.Journal

Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nature Reviews. Genetics, 7, 552–564. doi:10.1038/nrg1895

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V, Schwartz, S., … Eichler, E. E. (2002). Recent segmental duplications in the human genome. Science (New York, N.Y.), 297, 1003–1007. doi:10.1126/science.1072047

Bell, J. T., Tsai, P.-C., Yang, T.-P., Pidsley, R., Nisbet, J., Glass, D., … Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and

age-related phenotypes in a healthy ageing population. PLoS Genetics, 8(4), e1002629. doi:10.1371/journal.pgen.1002629

Benros, M. E., Mortensen, P. B., & Eaton, W. W. (2012). Autoimmune diseases and infections as risk factors for schizophrenia. Annals of the New York Academy of Sciences, 1262, 56–66. doi:10.1111/j.1749-6632.2012.06638.x

Berg, J. S., Brunetti-Pierri, N., Peters, S. U., Kang, S.-H. L., Fong, C., Salamone, J., … Cheung, S. W. Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. , 9 Genetics in medicine : official journal of the American College of Medical Genetics 427–441 (2007). doi:10.1097/GIM.0b013e3180986192

Bergé, D., Carmona, S., Rovira, M., Bulbena, a, Salgado, P., & Vilarroya, O. (2011). Gray matter volume deficits and correlation with insight and negative symptoms in first-psychotic-episode subjects. Acta Psychiatrica Scandinavica, 123(6), 431–9. doi:10.1111/j.1600-0447.2010.01635.x

Bergen, S. E., & Petryshen, T. L. (2012). Genome-wide association studies of schizophrenia: does bigger lead to better results? Current Opinion in Psychiatry, 25(2), 76–82. doi:10.1097/YCO.0b013e32835035dd

Blackledge, N. P., & Klose, R. J. (2011). CpG island chromatin: A platform for gene regulation. Epigenetics, 6, 147–152. doi:10.4161/epi.6.2.13640

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics, 40, 695–701. doi:10.1038/ng.f.136

Bondeson, M. L., Dahl, N., Malmgren, H., Kleijer, W. J., Tönnesen, T., Carlberg, B. M., & Pettersson, U. (1995). Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. Human Molecular Genetics, 4, 615–621. doi:10.1093/hmg/4.4.615

Bosch, N., Morell, M., Ponsa, I., Mercader, J. M., Armengol, L., & Estivill, X. (2009). Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. PLoS ONE, 4. doi:10.1371/journal.pone.0008269

Bowness, P. (2002). HLA B27 in health and disease: a double-edged sword? Rheumatology (Oxford, England), 41(8), 857–68. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12154202

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., … Rizzolatti, G. (2004). Neural Circuits Involved in the Recognition of Actions Performed by Nonconspecifics: An fMRI Study, 114–126.

Bucher, E., Reinders, J., & Mirouze, M. (2012). Epigenetic control of transposon transcription and mobility in Arabidopsis. Current Opinion in Plant Biology, 15(5), 503–10. doi:10.1016/j.pbi.2012.08.006

Burbach, J. P. H., & Zwaag, B. Van Der. (2009). Contact in the genetics of autism and schizophrenia. Cell, 15–18.

Cáceres, A., & González, J. R. (2015). Following the footprints of polymorphic inversions on SNP data: from detection to association tests. Nucleic Acids Research, 1–11. doi:10.1093/nar/gkv073

Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M., & González, J. R. (2012). Identification of polymorphic inversions from genotypes. BMC Bioinformatics. doi:10.1186/1471-2105-13-28

Campbell, I. M., Shaw, C. a, Stankiewicz, P., & Lupski, J. R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends in Genetics : TIG, 1–11. doi:10.1016/j.tig.2015.03.013

Carlini, D. B., & Stephan, W. (2003). In vivo introduction of unpreferred synonymous codons into the drosophila Adh gene results in reduced levels of ADH protein. Genetics, 163, 239–243.

Carroll, L. S., & Owen, M. J. (2009). Genetic overlap between autism, schizophrenia and bipolar disorder. Genome Medicine, 1, 102. doi:10.1186/gm102

Castillejo-Lopez, C., Delgado-Vega, A. M., Wojcik, J., Kozyrev, S. V., Thavathiru, E., Wu, Y.-Y., … Alarcon-Riquelme, M. E. (2012). Genetic and physical interaction of the B-cell systemic lupus erythematosus-associated genes BANK1 and BLK. Annals of the Rheumatic Diseases. doi:10.1136/annrheumdis-2011-200085

Centers for Disease Control and Prevention. (2012). Prevalence of autism spectrum disorders autism and developmental disabilities monitoring network, 14 sites, United States, 2008. Morbidity and mortality weekly report. Surveillance summaries. Centers for Disease Control and Prevention, 61, 1–19.

Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L.-C., & Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biology, 4, R25. doi:10.1186/gb-2003-4-4-r25

Chien, H.-Y., Gau, S. S.-F., Hsu, Y.-C., Chen, Y.-J., Lo, Y.-C., Shih, Y.-C., & Tseng, W.-Y. I. (2015). Altered Cortical Thickness and Tract Integrity of the Mirror Neuron System and Associated Social Communication in Autism Spectrum Disorder. Autism Research, n/a–n/a. doi:10.1002/aur.1484

Chisholm, K., Lin, A., Abu-Akel, A., & Wood, S. J. (2015). The association between autism and schizophrenia spectrum disorders: A review of eight alternate models of co-occurrence. Neuroscience & Biobehavioral Reviews, 55, 173–183. doi:10.1016/j.neubiorev.2015.04.012

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews. Genetics, 11, 415–425. doi:10.1038/nrg2779

Codina-Solà, M., Rodríguez-Santiago, B., Homs, A., Santoyo, J., Rigau, M., Aznar-Laín, G., … Cuscó, I. (2015). Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. Molecular Autism, 6(1), 21. doi:10.1186/s13229-015-0017-0

Colitis, U. (2003). Inflammatory Bowel Disease Part I: Ulcerative Colitis – Pathophysiology and Conventional and Alternative Treatment Options, 8(3).

Cooke, M. a, Fannon, D., Kuipers, E., Peters, E., Williams, S. C., & Kumari, V. (2008). Neurological basis of poor insight in psychosis: a voxel-based MRI study. Schizophrenia Research, 103(1-3), 40–51. doi:10.1016/j.schres.2008.04.022

Cowley, M., & Oakey, R. J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. PLoS Genetics. doi:10.1371/journal.pgen.1003234

Coyne, J. A., Aulard, S., & Berry, A. (1991). Lack of underdominance in a naturally occurring pericentric inversion in drosophila melanogaster and its implications for chromosome evolution. Genetics, 129, 791–802.

Crespi, B., & Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain, 241–320.

Daley, J. M., Palmbos, P. L., Wu, D., & Wilson, T. E. (2005). Nonhomologous end joining in yeast. Annual Review of Genetics, 39, 431–451. doi:10.1146/annurev.genet.39.073003.113340

Danese, S., & Fiocchi, C. (2011). Ulcerative Colitis. The New England Journal O F Medicine - Review, 1713–1725.

De Bakker, P. I. W., & Raychaudhuri, S. (2012). Interrogating the major histocompatibility complex with high-throughput genomics. Human Molecular Genetics, 21(R1), R29–36. doi:10.1093/hmg/dds384

De Fossé, L., Hodge, S. M., Makris, N., Kennedy, D. N., Caviness, V. S., McGrath, L., … Harris, G. J. (2004). Language-association cortex asymmetry in autism and specific language impairment. Annals of Neurology, 56(6), 757–66. doi:10.1002/ana.20275

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. Genes & Development, 25, 1010–1022. doi:10.1101/gad.2037511

Dennis, M. K., Field, A. S., Burai, R., Ramesh, C., Whitney, K., Bologa, C. G., … Prossnitz, E. R. (2012). NIH Public Access, 127(9), 358–366. doi:10.1016/j.jsbmb.2011.07.002.Identification

Devlin, B., Melhem, N., & Roeder, K. (2011). Do common variants play a role in risk for autism? Evidence and theoretical musings. Brain Research, 1380, 78–84. doi:10.1016/j.brainres.2010.11.026

Ecker, J. R. (n.d.). News & views. 2012, 6–9.

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics : TIG, 24(7), 344–52. doi:10.1016/j.tig.2008.04.005

Engelken, J., Carnero-Montoro, E., Pybus, M., Andrews, G. K., Lalueza-Fox, C., Comas, D., … Bosch, E. (2014). Extreme Population Differences in the Human Zinc Transporter ZIP4 (SLC39A4) Are Explained by Positive Selection in Sub-Saharan Africa. PLoS Genetics, 10. doi:10.1371/journal.pgen.1004128

Esteller, M. (2011). Non-coding RNAs in human disease. Nature Reviews. Genetics, 12(12), 861–74. doi:10.1038/nrg3074

Fernando, M. M. a, Stevens, C. R., Walsh, E. C., De Jager, P. L., Goyette, P., Plenge, R. M., … Rioux, J. D. (2008). Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genetics, 4(4), e1000024. doi:10.1371/journal.pgen.1000024

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews. Genetics, 7(2), 85–97. doi:10.1038/nrg1767

Flores, R., Gutiérrez-Arumí, A., Verma, B., Oghabian, A., Chowen, J. A., & Frilander, M. J. (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency, 1–8.

Francke, U. (1999). Williams-Beuren syndrome: Genes and mechanisms. Human Molecular Genetics. doi:10.1093/hmg/8.10.1947

Fu, C. H. Y., Suckling, J., Williams, S. C. R., Andrew, C. M., Vythelingum, G. N., & Mcguire, P. K. (n.d.). Effects of Psychotic State and Task Demand on Prefrontal Function in Schizophrenia: An fMRI Study of Overt Verbal Fluency, 485–494.

Fujihara, K., Miwa, H., Kakizaki, T., Kaneko, R., Mikuni, M., Tanahira, C., … Yanagawa, Y. (2015). Glutamate Decarboxylase 67 Deficiency in a Subset of GABAergic Neurons Induces Schizophrenia-Related Phenotypes. Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology, 1–12. doi:10.1038/npp.2015.117

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. a, Goldberg, A. P., Lee, A. B., … Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. Nature Genetics, 46(8), 881–885. doi:10.1038/ng.3039

Gentilucci, M., Benuzzi, F., Bertolani, L., Daprati, E., & Gangitano, M. (2000). Language and motor control. Experimental Brain Research, 133(4), 468–490. doi:10.1007/s002210000431

Giglio, S., Broman, K. W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., … Zuffardi, O. (2001). Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. American Journal of Human Genetics, 68, 874–883. doi:10.1086/319506

Glessner, J. T., Li, J., & Hakonarson, H. (2013). ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Research, 41. doi:10.1093/nar/gks1346

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell, 128(4), 635–8. doi:10.1016/j.cell.2007.02.006

González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., … Pérez-Jurado, L. a. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. American Journal of Human Genetics, 94(3), 361–72. doi:10.1016/j.ajhg.2014.01.015

Gonzalez-Burgos, G., & Lewis, D. a. (2008). GABA neurons and the mechanisms of network oscillations: implications for understanding cortical dysfunction in schizophrenia. Schizophrenia Bulletin, 34(5), 944–61. doi:10.1093/schbul/sbn070

Gutierrez, R. C., Hung, J., Zhang, Y., Kertesz, a C., Espina, F. J., & Colicos, M. a. (2009). Altered synchrony and connectivity in neuronal networks expressing an autism-related mutation of neuroligin 3. Neuroscience, 162(1), 208–21. doi:10.1016/j.neuroscience.2009.04.062

Guttenbach, M., Koschorz, B., Bernthaler, U., Grimm, T., & Schmid, M. (1995). Sex chromosome loss and aging: in situ hybridization studies on human interphase nuclei. The American Journal of Human Genetics, 57, 1143–1150. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=7485166&retmode=ref&cmd=prlinks\npapers2://publication/uuid/D3EC1793-B184-4DD5-8356-9C6551909211

Häfner, H. (2003). Gender differences in schizophrenia. Psychoneuroendocrinology, 28, 17–54. doi:10.1016/S0306-4530(02)00125-7

Hale, C. M., & Tager-flusberg, H. (2003). The influence of language on theory of mind: a training study, 3, 346–359.

Hamilton, A. F. D. C., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: how valid is the hypothesis of a deficit in the mirror neuron system? Neuropsychologia, 45(8), 1859–68. doi:10.1016/j.neuropsychologia.2006.11.022

Hamlyn, J., Duhig, M., McGrath, J., & Scott, J. (2013). Modifiable risk factors for schizophrenia and autism - Shared risk factors impacting on brain development. Neurobiology of Disease. doi:10.1016/j.nbd.2012.10.023

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. Nature Reviews. Genetics, 10, 551–564. doi:10.1038/nrg2593

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews. Genetics, 5, 522–531. doi:10.1038/nrg1415

Holliday, R. (2014). Epigenetics: A Historical Overview. Epigenetics, 1(2), 76–80. doi:10.4161/epi.1.2.2762

Hommer, R. E., & Swedo, S. E. (2015). Schizophrenia and Autism-Related Disorders. Schizophrenia Bulletin, 10–11. doi:10.1093/schbul/sbu188

Hurles, M. E., & Lupski, J. R. (2006). Recombination hotspots in nonallelic homologous recombination. In Genomic Disorders: The Genomic Basis of Disease (pp. 341–355). doi:10.1007/978-1-59745-039-3_24

IAN R. M ACKAY, M. ., & FRED S. R OSEN, M. D. (2001). Autoimmune Diseases - Review. New England Journal of Human Genetics, 345(5), 340–350.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., … Wigler, M. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron, 74, 285–299. doi:10.1016/j.neuron.2012.04.009

Jacobs, K. B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., … Chanock, S. J. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nature Genetics. doi:10.1038/ng.2270

JACOBS, P. A., BAIKIE, A. G., COURT BROWN, W. M., & STRONG, J. A. (1959). The somatic chromosomes in mongolism. Lancet, 1, 710. doi:10.1016/S0140-6736(59)92207-X

Jansen, J., Karges, W., & Rink, L. (2009). Zinc and diabetes - clinical links and molecular mechanisms. Journal of Nutritional Biochemistry. doi:10.1016/j.jnutbio.2009.01.009

Just, M. A., Cherkassky, V. L., Keller, T. a, & Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. Brain : A Journal of Neurology, 127(Pt 8), 1811–21. doi:10.1093/brain/awh199

Kanduri, C., Kantojärvi, K., Salo, P. M., Vanhala, R., Buck, G., Blancher, C., … Järvelä, I. (2015). The landscape of copy number variations in Finnish families with autism spectrum disorders. Autism Research : Official Journal of the International Society for Autism Research, 1–8. doi:10.1002/aur.1502

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., … Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell, 143, 837–847. doi:10.1016/j.cell.2010.10.027

Kim, Y. S., & Leventhal, B. L. (2015). Genetic epidemiology and insights into interactive genetic and environmental effects in autism spectrum disorders. Biological Psychiatry, 77(1), 66–74. doi:10.1016/j.biopsych.2014.11.001

King, D. a, Jones, W. D., Crow, Y. J., Dominiczak, A. F., Foster, N. a, Gaunt, T. R., … Hurles, M. E. (2015). Mosaic structural variation in children with developmental disorders. Human Molecular Genetics, 24(10), 2733–2745. doi:10.1093/hmg/ddv033

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. Genetics, 173(1), 419–34. doi:10.1534/genetics.105.047985

Kitsios, G. (2010). Genome-Wide Association Studies: hypothesis-"free" or "engaged," 154(4), 161–164. doi:10.1016/j.trsl.2009.07.001.Genome-Wide

Klei, L., Sanders, S. J., Murtha, M. T., Hus, V., Lowe, J. K., Willsey, a J., … Devlin, B. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Molecular Autism, 3(1), 9. doi:10.1186/2040-2392-3-9

Koolen, D. a, Kramer, J. M., Neveling, K., Nillesen, W. M., Moore-Barton, H. L., Elmslie, F. V, … de Vries, B. B. a. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nature Genetics, 44(6), 639–41. doi:10.1038/ng.2262

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., … Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science (New York, N.Y.), 318, 420–426. doi:10.1126/science.1149504

Kosoy, R., Nassir, R., Tian, C., White, P. a, Butler, L. M., Silva, G., … Seldin, M. F. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human Mutation, 30(1), 69–78. doi:10.1002/humu.20822

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Research, 37(13), 4181–93. doi:10.1093/nar/gkp552

Lakich, D., Kazazian, H. H., Antonarakis, S. E., & Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nature Genetics, 5, 236–241. doi:10.1038/ng1193-236

Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., … Gerstein, M. B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nature Biotechnology, 28, 47–55. doi:10.1038/nbt.1600

Lieber, M. R. (2008). The mechanism of human nonhomologous DNA End joining. Journal of Biological Chemistry. doi:10.1074/jbc.R700039200

Lin, G. N., Corominas, R., Lemmens, I., Yang, X., Tavernier, J., Hill, D. E., … Iakoucheva, L. M. (2015). Spatiotemporal 16p11.2 Protein Network Implicates Cortical Late Mid-Fetal Brain Development and KCTD13-Cul3-RhoA Pathway in Psychiatric Diseases. Neuron, 85(4), 742–754. doi:10.1016/j.neuron.2015.01.010

Ma, J., Xiong, M., You, M., Lozano, G., & Amos, C. I. (2014). Genome-wide association tests of inversions with application to psoriasis. Human Genetics, 133(8), 967–974. doi:10.1007/s00439-014-1437-1

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids Research, 42. doi:10.1093/nar/gkt958

Macdonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., … Harper, P. S. (1993). A Novel Gene Containing A Trinucleotide Repeat That Is Expanded and Unstable on Huntingtons-Disease Chromosomes. Cell, 72, 971–983. doi:10.1016/0092-8674(93)90585-E

Maher, B. (2008). The case of missing heritability. Nature Reviews. Genetics.

Manolio, T. a, Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. a, Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–53. doi:10.1038/nature08494

Manolio, T. A. (2009). Cohort studies and the genetics of complex disease, 41(1), 5–6.

Marco, E. J., & Skuse, D. H. (2006). Autism-lessons from the X chromosome. Social Cognitive and Affective Neuroscience. doi:10.1093/scan/nsl028

Martens, M. A., Wilson, S. J., & Reutens, D. C. (2008). Research Review: Williams syndrome: A critical review of the cognitive, behavioral, and neuroanatomical phenotype. Journal of Child Psychology and Psychiatry and Allied Disciplines. doi:10.1111/j.1469-7610.2008.01887.x

Mazefsky, C. A., & Oswald, D. P. (2006). The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting. Autism : The International Journal of Research and Practice, 10, 533–549. doi:10.1177/1362361306068505

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. a, & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews. Genetics, 9(5), 356–69. doi:10.1038/nrg2344

McConnell, M. J., Lindberg, M. ., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., … Gage, F. H. (2013). Mosaic Copy Number Variation in Human Neurons. Science, 342, 632–638. doi:10.1126/science.1243472

McGovern, C. W., & Sigman, M. (2005). Continuity and change from early childhood to adolescence in autism. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 46(4), 401–8. doi:10.1111/j.1469-7610.2004.00361.x

Mead, J., & Ashwood, P. (2015). Evidence supporting an altered immune response in ASD. Immunology Letters, 163(1), 49–55. doi:10.1016/j.imlet.2014.11.006

Meyer, U. R. S., Feldon, J., & Dammann, O. (2011). Schizophrenia and Autism: Both Shared and Disorder-Specific Pathogenesis. Pediatric Research, 69(5), 26–33.

MHC sequencing consortium. (1999). letters to nature. Nature, 401(October), 921–923.

Miles, J. H. (2011). Autism spectrum disorders--a genetics review. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 13(4), 278–94. doi:10.1097/GIM.0b013e3181ff67ba

Morris, K. V. (2014a). Long antisense non-coding RNAs function to direct epigenetic complexes that regulate transcription in human cells. Epigenetics, 4(5), 296–301. doi:10.4161/epi.4.5.9282

Morris, K. V. (2014b). Non-coding RNAs, epigenetic memory and the passage of information to progeny. RNA Biology, 6(3), 242–247. doi:10.4161/rna.6.3.8353

Muotri, A. R. (2015). The human model: changing focus on autism research. Biological Psychiatry, 1–8. doi:10.1016/j.biopsych.2015.03.012

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., … Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. doi:10.1038/nature11011

Nesse, R. M., & Williams, G. C. (1998). Evolution and the origins of disease. Scientific American, 279, 86–93. doi:10.1038/scientificamerican1198-86

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., … Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. doi:10.1038/nature10989

Oberman, L. M., Ramachandran, V. S., & Pineda, J. a. (2008). Modulation of mu suppression in children with autism spectrum disorders in response to familiar or unfamiliar stimuli: the mirror neuron hypothesis. Neuropsychologia, 46(5), 1558–65. doi:10.1016/j.neuropsychologia.2008.01.010

Ohno, S. (1972). So much "junk" DNA in our genome.

Ornoy, a, Weinstein-Fudim, L., & Ergaz, Z. (2015). Prenatal factors associated with Autism Spectrum Disorder (ASD). Reproductive Toxicology (Elmsford, N.Y.). doi:10.1016/j.reprotox.2015.05.007

Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., … Scherer, S. W. (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nature Genetics, 29, 321–325. doi:10.1038/ng753

Pagani, F., Raponi, M., & Baralle, F. E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proceedings of the National Academy of Sciences of the United States of America, 102, 6368–6372. doi:10.1073/pnas.0502288102

Pang, A. W. C., Migita, O., Macdonald, J. R., Feuk, L., & Scherer, S. W. (2013). Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. Human Mutation, 34, 345–354. doi:10.1002/humu.22240

Perra, O., Williams, J. H. G., Whiten, A., Fraser, L., Benzie, H., & Perrett, D. I. (2008). Imitation and "theory of mind" competencies in discrimination of autism from other neurodevelopmental disorders. Research in Autism Spectrum Disorders, 2(3), 456–468. doi:10.1016/j.rasd.2007.09.007

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., … Scherer, S. W. (2014). Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. The American Journal of Human Genetics, 677–694. doi:10.1016/j.ajhg.2014.03.018

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., … Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature, 466(7304), 368–72. doi:10.1038/nature09146

Piotrowski, A., Bruder, C. E. G., Andersson, R., De Ståhl, T. D., Menzel, U., Sandgren, J., … Dumanski, J. P. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. Human Mutation, 29, 1118–1124. doi:10.1002/humu.20815

Pique-Regi, R., Cáceres, A., & González, J. R. (2010). R-Gada: a fast and flexible pipeline for copy number analysis in association studies. BMC Bioinformatics, 11, 380. doi:10.1186/1471-2105-11-380

Plomin, R., & Deary, I. J. (2014). Genetics and intelligence differences: five special findings. Molecular Psychiatry, (July 2014), 98–108. doi:10.1038/mp.2014.105

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., & Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 465, 1033–1038. doi:10.1038/nature09144

Pramparo, T., Pierce, K., Lombardo, M. V., Carter Barnes, C., Marinero, S., Ahrens-Barbeau, C., … Courchesne, E. (2015). Prediction of Autism by Translation and Immune/Inflammation Coexpressed Genes in Toddlers From Pediatric Community Practices. JAMA Psychiatry, 92093(4), 386–394. doi:10.1001/jamapsychiatry.2014.3008

Qumsiyeh, M. B., Kim, K. R., Ahmed, M. N., & Bradford, W. (2000). Cytogenetics and mechanisms of spontaneous abortions: increased apoptosis and decreased cell proliferation in chromosomally abnormal villi. Cytogenetics and Cell Genetics, 88, 230–235. doi:10.1159/000015557

Radke, D. W., & Lee, C. (2015). Adaptive potential of genomic structural variation in human and mammalian evolution. Briefings in Functional Genomics, 1–11. doi:10.1093/bfgp/elv019

Raymond, L. A., André, V. M., Cepeda, C., Gladding, C. M., Milnerwood, A. J., & Levine, M. S. (2011). Pathophysiology of Huntington's disease: Time-dependent alterations in synaptic and receptor function. Neuroscience. doi:10.1016/j.neuroscience.2011.08.052

Report, M. W. (2014). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C. : 2002), 63, 1–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24670961

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. a., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature. doi:10.1038/nature13595

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions, 3, 131–141.

Rodríguez-Santiago, B., Malats, N., Rothman, N., Armengol, L., Garcia-Closas, M., Kogevinas, M., … Pérez-Jurado, L. A. (2010). Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. American Journal of Human Genetics, 87, 129–138. doi:10.1016/j.ajhg.2010.06.002

Rosin, B., Slovik, M., Mitelman, R., Rivlin-Etzion, M., Haber, S. N., Israel, Z., … Bergman, H. (2011). Closed-loop deep brain stimulation is superior in ameliorating parkinsonism. Neuron, 72, 370–384. doi:10.1016/j.neuron.2011.08.023

Ross, M. T., Grafham, D. V, Coffey, A. J., Scherer, S., McLay, K., Muzny, D., … Bentley, D. R. (2005). The DNA sequence of the human X chromosome. Nature, 434, 325–337. doi:10.1038/nature03440

Rossignol, D. a, & Frye, R. E. (2012). A review of research trends in physiological abnormalities in autism spectrum disorders: immune dysregulation, inflammation, oxidative stress, mitochondrial dysfunction and environmental toxicant exposures. Molecular Psychiatry, 17(4), 389–401. doi:10.1038/mp.2011.165

Russo, C., Carroll, A., Kohler, S., Borowitz, M., Amylon, M., Homans, A., … Crist, W. (1991). Philadelphia chromosome and monosomy 7 in childhood acute lymphoblastic leukemia: a Pediatric Oncology Group study. Blood, 77, 1050–1056.

Salm, M. P. a, Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., … Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Research, 22(6), 1144–1153. doi:10.1101/gr.126037.111

Samuel P. Strom, Jennifer L. Stone, John R. ten Bosch, Barry Merriman, R. M., & Cantor, Daniel H. Geschwind, and S. F. N. (2011). High Density SNP Association Study of the 17q21 Chromosomal Region Linked to Autism Identifies CACNA1G as a Novel Candidate Gene, 15(10), 996–1005. doi:10.1038/mp.2009.41.

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., … State, M. W. (2011). Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron, 70, 863–885. doi:10.1016/j.neuron.2011.05.002

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., … State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. doi:10.1038/nature10945

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-martin, C., Walsh, T., … Wigler, M. (2007). Strong Association of De Novo Copy Number Mutations with Autism, 316(April), 445–449.

Seeman, P., Tedesco, J., & Wong, K. (1975). antipsychotic drugs, 72(11), 4376–4380.

Shad, M. U., & Keshavan, M. S. (2015). Neurobiology of insight deficits in schizophrenia: An fMRI study. Schizophrenia Research, 165(2-3), 220–226. doi:10.1016/j.schres.2015.04.021

Shao, L., Shaw, C. A., Lu, X. Y., Sahoo, T., Bacino, C. A., Lalani, S. R., … Cheung, S. W. (2008). Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: A study of 5,380 cases. American Journal of Medical Genetics, Part A, 146, 2242–2251. doi:10.1002/ajmg.a.32399

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135–45. doi:10.1038/nbt1486

Sherry, S. T., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation, 29(1), 308–311.

Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., … Gejman, P. V. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature, 460, 753–757. doi:10.1038/nature08192

Simard, F., Ayala, D., Kamdem, G. C., Pombi, M., Etouna, J., Ose, K., … Costantini, C. (2009). Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. BMC Ecology, 9, 17. doi:10.1186/1472-6785-9-17

Somerville, M. J., Mervis, C. B., Young, E. J., Seo, E., Bamforth, S., Peregrine, E., … Osborne, L. R. (2015). Severe Expressive-Language Delay Related to Duplication of the Williams–Beuren Locus, 1694–1701.

Spek, A. A., & Wouters, S. G. M. (2010). Autism and schizophrenia in high functioning adults: Behavioral differences and overlap. Research in Autism Spectrum Disorders, 4, 709–717. doi:10.1016/j.rasd.2010.01.009

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., … Stefansson, K. (2005). A common inversion under selection in Europeans. Nature Genetics, 37, 129–137. doi:10.1038/ng1508

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., … Collier, D. A. (2009). Common variants conferring risk of schizophrenia. Nature, 460, 744–747. doi:10.1038/nature08186

Sturtevant, A. (1921). A CASE OF REARRANGEMENT OF GENES IN DROSOPHILA. Genetics, 7, 235–237.

Tan, C. C. (1935). Salivary gland chromosomes in the two races of Drosophila pseudoobscura., 392–402.

Trowsdale, J., & Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. Annual Review of Genomics and Human Genetics, 14, 301–23. doi:10.1146/annurev-genom-091212-153455

Tseng, H.-H., Bossong, M. G., Modinos, G., Chen, K.-M., McGuire, P., & Allen, P. (2015). A systematic review of multisensory cognitive-affective integration in schizophrenia. Neuroscience and Biobehavioral Reviews, 1–9. doi:10.1016/j.neubiorev.2015.04.019

Turner, D. J., Shendure, J., Porreca, G., Church, G., Green, P., Tyler-smith, C., & Hurles, M. E. (2006). Assaying chromosomal inversions by single-molecule haplotyping. Nature Communications, 3(6), 439–445. doi:10.1038/NMETH881

Turner, T. N., Sharma, K., Oh, E. C., Liu, Y. P., Collins, R. L., Sosa, M. X., … Chakravarti, A. (2015). Loss of δ-catenin function in severe autism. Nature. doi:10.1038/nature14186

Uhlhaas, P. J., & Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. Nature Reviews. Neuroscience, 11(2), 100–13. doi:10.1038/nrn2774

Van der Aa, N., Rooms, L., Vandeweyer, G., van den Ende, J., Reyniers, E., Fichera, M., … Kooy, R. F. (2009). Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. European Journal of Medical Genetics, 52, 94–100. doi:10.1016/j.ejmg.2009.02.006

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution, (22).

Varki, A., & Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack, 1746–1758. doi:10.1101/gr.3737405.evolutionary

Veenstra-VanderWeele, J., & Blakely, R. D. (2012). Networking in autism: leveraging genetic, biomarker and model system findings in the search for new treatments. Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology, 37(1), 196–212. doi:10.1038/npp.2011.185

Vernes, S. C., Newbury, D. F., Abrahams, B. S., Winchester, L., Groszer, M., Oliver, P. L., … Fisher, S. E. (2015). A Functional Genetic Link between Distinct Developmental Language Disorders. New England Journal of Human Genetics.

Viguera, E., Canceill, D., & Ehrlich, S. D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. EMBO Journal, 20, 2587–2595. doi:10.1093/emboj/20.10.2587

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era--concepts and misconceptions. Nature Reviews. Genetics, 9(4), 255–266. doi:10.1038/nrg2322

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., … Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research, 17, 1665–1674. doi:10.1101/gr.6861907

Wang, Y., Peoples, R., Coloma, A., & Francke, U. (1998). A duplicated gene in the breakpoint regions of the 7q11.23 Williams–Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK, 7(3), 325–334.

Williams, G. C. (1957). Pleiotropy, Natural Selection, and the Evolution of Senescence. Evolution, 11, 398–411. doi:10.2307/2406060

Wray, N. R., Goddard, M. E., & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies, 1520–1528. doi:10.1101/gr.6665407.1520

Yang, D. Y.-J., Rosenblau, G., Keifer, C., & Pelphrey, K. a. (2015). An integrative neural model of social perception, action observation, and theory of mind. Neuroscience and Biobehavioral Reviews, 51, 263–275. doi:10.1016/j.neubiorev.2015.01.020

Zollino, M., Orteschi, D., Murdolo, M., Lattante, S., Battaglia, D., Stefanini, C., … Marangi, G. (2012). Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. Nature Genetics, 44(6), 636–8. doi:10.1038/ng.2257

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1119675109

9. APPENDIX

Mendelian disorder produced by minor spliceosome mRNA processing

Background: The molecular basis of a significant number of cases of isolated growth hormone deficiency remains unknown. This is the first report of the implication of a minor spliceosome mutation in human disease. Specifically, isolated growth hormone deficiency due to a pituitary development defect affecting somatotrophs can derive from aberrant RNA splicing by RNPC3. We describe three sisters affected with severe isolated growth hormone deficiency and pituitary hypoplasia caused by biallelic mutations in the RNPC3 gene, which codes for a minor spliceosome protein required for U11/U12 small nuclear ribonucleoprotein (snRNP) formation and splicing of U12-type introns. Isolated growth hormone deficiency can be caused by aberrant RNA splicing by the minor spliceosome.

Methods: We performed blood cell transcriptome by RNAseq in the samples, and compared against a subset of controls in order to identify differences in splicing of alternative spliceing using in-house methods. Transcriptomic analysis allowed the identification of putative genes that could explain the pituitary hypoplasia.

Results: We found anomalies in U11/U12 di-snRNP formation and in splicing of multiple U12-type introns in patient cells. Defective transcripts include preprohormone convertases SPCS2 and SPCS3 and actin-related ARPC5L genes, which are candidates for the somatotroph-restricted dysfunction. The reported novel mechanism for familial growth hormone deficiency demonstrates that general mRNA processing defects of the minor spliceosome can lead to very narrow tissue-specific consequences.

Jesús Argente, Raquel Flores, Armand Gutiérrez-Arumí, Bhupendra Verma, Gabriel A. Martos-Moreno, Ivon Cuscó, Ali Oghabian, Julie A Chowen, Mikko J Frilander, Luis A Pérez-Jurado (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. EMBO Molecular Medicine DOI 10.1002/emmm.201303573

Integrated analysis of whole-exome sequencing and transcriptome on ASD

Background: Autism spectrum disorders (ASD) are a group of neurodevelopmental disorders with high heritability. Recent findings support a highly heterogeneous and complex genetic etiology including rare de novo and inherited mutations or chromosomal rearrangements as well as double or multiple hits.

Methods: We performed whole-exome sequencing (WES) and blood cell transcriptome by RNAseq in a subset of male patients with idiopathic ASD (n = 36) in order to identify causative genes, transcriptomic alterations, and susceptibility variants.
Results: We detected likely monogenic causes in seven cases: five de novo (SCN2A, MED13L, KCNV1, CUL3, and PTEN) and two inherited X-linked variants (MAOA and CDKL5). Transcriptomic analyses allowed the identification of intronic causative mutations missed by the usual filtering of WES and revealed functional consequences of some rare mutations. These included aberrant transcripts (PTEN, POLR3C), deregulated expression in 1.7% of mutated genes (that is, SEMA6B, MECP2, ANK3, CREBBP), allele-specific expression (FUS, MTOR, TAF1C), and non-sense-mediated decay (RIT1, ALG9). The analysis of rare inherited variants showed enrichment in relevant pathways such as the PI3K-Akt signaling and the axon guidance.

Conclusions: Integrative analysis of WES and blood RNAseq data has proven to be an efficient strategy to identify likely monogenic forms of ASD (19% in our cohort), as well as additional rare inherited mutations that can contribute to ASD risk in a multifactorial manner. Blood transcriptomic data, besides validating 88% of expressed variants, allowed the identification of missed intronic mutations and revealed functional correlations of genetic variants, including changes in splicing, expression levels, and allelic expression.

Marta Codina-Solà, Benjamín Rodríguez-Santiago, Aïda Homs, Javier Santoyo, Maria Rigau, Gemma Aznar-Laín, Miguel del Campo, Blanca Gener, Elisabeth Gabau, María Pilar Botella, Armand Gutiérrez-Arumí, Guillermo Antiñolo, Luis Alberto Pérez-Jurado, Ivon Cuscó (2015). Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disordersMolecular Autism, 6(1), 21. doi:10.1186/s13229-015-0017-0