The value of studying the Indian population to identify novel genetic variants to inform mechanisms of disease and pharmacological response

A. Das[1] , P. Raj[2], V. Gopalan[2], Hiranjith. G.H[1] , E. Stawiski [1], S. Santhosh[1], R. Gupta[2] , A. Chaudhuri[1] ,R. Gupta[2]

1) MedGenome Inc, Foster City, CA, 2) MedGenome Labs, Bangalore, India; corresponding author : ankita.d@medgenome.com

Abstract

While Genome wide association studies can shed light on the significance of variants in susceptibility to a disease or allow to stratify patients for specific therapeutic modalities, often variants that are rare and could be of significance are not identified in these studies. This can occur due to allelic heterogeneity in a complex disease. Furthermore, spurious differences in allelic frequencies between normal and disease resulting from systematic differences in ancestry can also confound the conclusions drawn from a GWAS study. Therefore, studying population isolates where individuals with the disease and normal have a homogeneous genetic background can allow to enrich for rare alleles, and improve the accuracy of elimination of false positives, and make it possible to accurately correlate segregation of the variants to the disease traits. One such population is of the Indian sub-continent, where the ancestral populations date back to modern humans travelling out of Africa 65,000 year ago, creating a gene pool of over 1000 years starting from a few founder families, resulting in an accumulation of unique disease-causing and disease-protective alleles that were preserved and enriched within various ethnic groups in the country. In addition to the rich genetic diversity, there are geographically isolated sub-populations that are relatively homogeneous genetically due to the endogamy practices. These sub-population isolates are ideal for enrichment of rare alleles that can identify genetic risk factors associated with certain diseases and will allow for stratified population analysis to understand and estimate the risk of disease related variants across sub-populations. Furthermore, the common complex diseases such as inherited diseases, diabetes, cancers are widely prevalent in certain sub-populations within the Indian population, making it a rich population to study to identify rare alleles and disease-causing variants. At MedGenome, we provide researchers with an integrated solution with access to specific cohorts with clinical diagnosis & familial history information, genome sequencing to enable genetics and pharmacogenomics research and analytics to interpret the data. We present case examples of utilizing the cohorts from the Indian population to find occurrence of cancer driver genes, and find variants of significance in muscular dystrophy and pharmacogenomics analyses. By examining genes that carry variants known to cause disease in the Indian cohort datasets, we identify minor allele frequency of the variants across populations and also identify novel variants in the genes that can be further studied to understand mechanisms of disease and identify potential biomarkers and novel drug targets.

The value of studying the Indian population

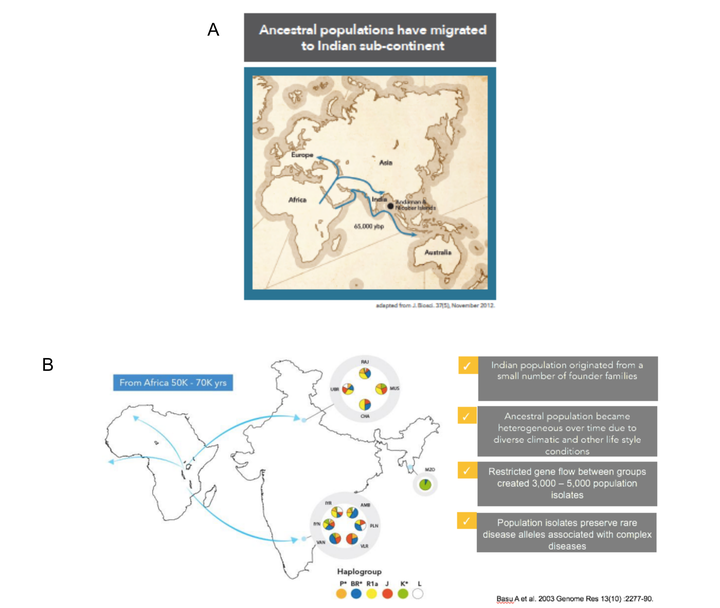


Figure 1: A. Shows the pattern of migration into the South-east Asian sub-continents showing that the ancestral populations migrated through the Indian subcontinent. B) show the Indian sub-continent is a rich genetic trove and the Indian population originated from a few founder families.

MedGenome can provide access to disease specific cohorts through its pan-India research network

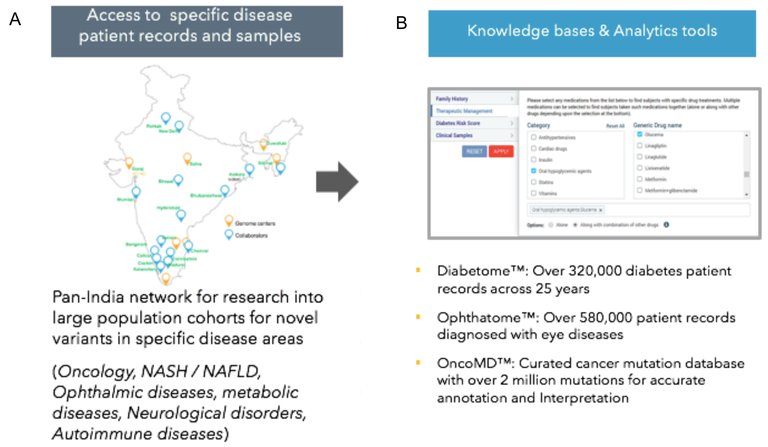
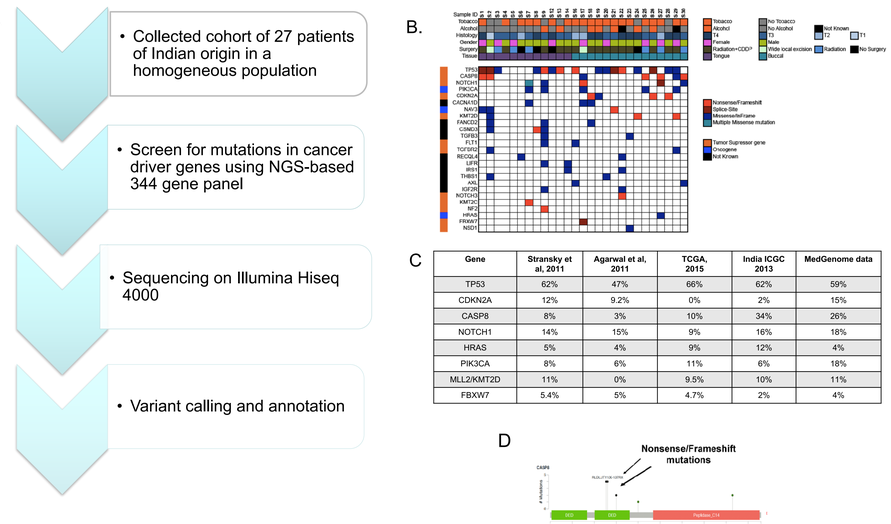


Figure 2: A. Shows the MedGenome network of research collaborations pan-India to gain access to disease-specific cohorts with well-characterized phenotype information for human genetics research and pharmacogenomic studies. B. Analytic tools by MedGenome to query for disease-specific cohorts by phenotype criteria for research

Analysis of a cohort of head and neck cancer patients of Indian origin identifies a high frequency of occurrence of mutation in driver gene Casp8



Gene	Stransky et al, 2011	Agarwal et al, 2011	TCGA, 2015	India ICGC 2013	MedGenome data
TP53	62%	47%	66%	62%	59%
CDKN2A	12%	9.2%	0%	2%	15%
CASP8	8%	3%	10%	34%	26%
NOTCH1	14%	15%	9%	16%	18%
HRAS	5%	4%	9%	12%	4%
PIK3CA	8%	6%	11%	6%	18%
MLL2/KMT2D	11%	0%	9.5%	10%	11%
FBXW7	5.4%	5%	4.7%	2%	4%

Figure 3: Screening and analysis of driver mutations in a cohort of 27 head and neck cancer patients of Indian origin : A) Workflow of collection and screening of a cohort of head and neck cancer patients to identify the frequency of mutations in cancer driver genes using an NGS-based 344 gene panel. B) Heatmap shows the distribution of the mutations in the different patients that were screened by this assay. C) Table shows the frequency of occurrence of the common driver genes across different databases that contain Caucasian versus Indian population data. Results show that the frequency of mutations are similar across different populations except Casp8, which has a high frequency of occurrence in the Indian population based on analysis of the MedGenome cohort and the ICGC databases. D) Shows a map of Casp8 functional domains and frequently occurring loss of function mutations in H&N cancer.

Screening for variants on SMN2 gene relevant for Duchenne muscular dystrophy in the MedGenome cohorts identifies known variants of significance and additional novel variants

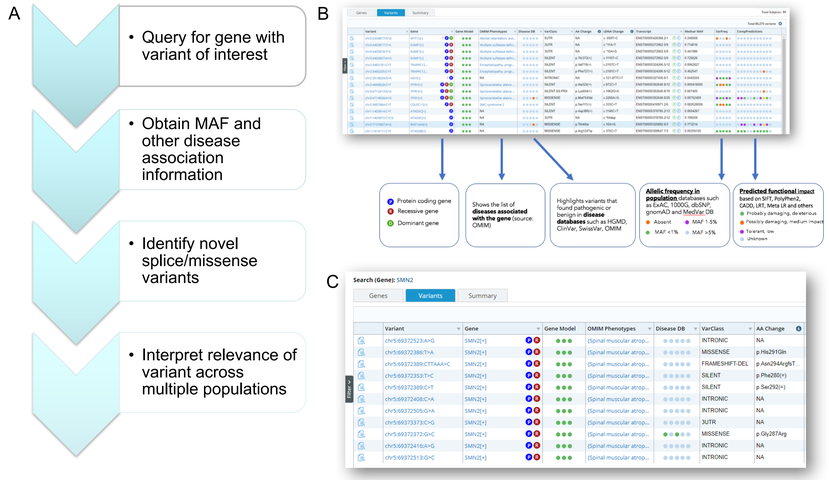
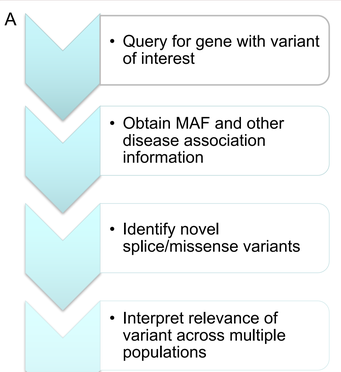


Figure 4: A) Workflow to analyze variants in SMN2 gene and identify the frequency of occurrence of the disease causing splice variants in the cohort of Indian samples. B) Overview of the types of output obtained upon analysis of genetic variants. C) Summary of the variants found in the SMN2 gene. The known variant for causing exon skipping and disease is captured and is present in the cohort of Indian population, and other novel splice/missense variants are identified that can be studied further to identify novel therapeutic targets and understand differential response to the splice blocker drug.

Studying disease-specific cohorts from India can provide insights into pharmacological responses



B

Gene	Position	Ref allele	Alt Allele	MedGenome cohort af	dbsnp_af	g1000_overall_af	exac_overall_af	gene_functional_impact
RYR1	38983180	G	T	0.157708	0.05491	0.0549121	0.0693702	MISSENSE
CYP4F2	15990431	C	T	0.416169	0.2368	0.236821	0.272576	MISSENSE
CYP4F2	15996828	G	C	0.0152524	0.004193	0.00419329	0.00624835	MISSENSE
CYP2A6	41349759	T	C	0.0145505	0.004792	0.00479233	0.0115543	MISSENSE
CYP2A6	41350664	A	T	0.874884	0.9864	0.986422	0.987624	MISSENSE
XRCC1	44048333	G	A	0.00166759	0.0001997	0.00019968	0.00048658	MISSENSE
CYP2B6	41515263	A	G	0.327533	NA	NA	0.0563167	MISSENSE
XRCC1	44056412	C	T	0.106101	0.06709	0.0670927	0.0881128	MISSENSE
APOE	45411941	T	C	0.087473	0.1506	0.150559	0.184332	MISSENSE
ERCC1	45911330	T	A	0.19684	0.2865	0.286542	NA	3UTR
ERCC1	45912060	A	C	0.257173	0.2867	0.286741	0.208573	3UTR
ERCC1	45912343	A	G	0.257606	0.2867	0.286741	0.208222	3UTR
ERCC1	45912406	G	A	0.257125	0.2825	0.282548	0.207133	3UTR
CYP4F2	15989733	T	C	0.173368	NA	NA	0.369331	MISSENSE
CYP2A6	41357777	C	T	0.0142449	0.005192	0.00519169	0.00458881	MISSENSE
CYP2B6	41512841	G	T	0.381769	0.3157	0.315695	0.27319	MISSENSE
CYP2A6	41349874	T	A	0.025459	NA	NA	0.102002	MISSENSE
CYP2B6	41522715	C	T	0.0778457	0.05351	0.0535144	0.090906	MISSENSE
XRCC1	44050749	G	T	2.69673E-05	NA	NA	NA	NONSENSE-SS-PRX
APOE	45412079	C	T	0.0245963	0.07508	0.0750799	0.0718191	MISSENSE
CYP4F2	15996820	G	A	0.235302	0.233	0.233027	0.286405	MISSENSE
ERCC1	45912488	CAAG	C	0.201148	NA	NA	0.236617	3UTR
CYP4F2	15608257	T	C	0.160682	0.1801	0.180112	0.168004	5UTR
CYP4F2	16008388	A	C	0.159923	0.1571	0.157149	0.160631	MISSENSE
CYP4F2	16008388	A	C	0.159923	0.1571	0.157149	0.160631	5UTR
IFNL3	39735106	T	C	0.217396	0.3654	0.365415	0.278663	MISSENSE
CYP4F2	15989696	G	C	0.0660421	NA	NA	0.0950882	MISSENSE
CYP2A6	41352802	G	A	0.000539578	0.0001997	0.00019968	0.00002479	MISSENSE

Figure 5: A) Workflow to analyze variants in pharmacogenes and identify the frequency of occurrence of variants of significance in the cohort of Indian samples. B) Table summarizing the variants identified to be present in the cohort of the Indian population studied and comparison of the occurrence of the variants across other population databases.

Summary and conclusions

- Indian population is a genetic trove, and can enable exploratory research & drug target discovery
- MedGenome has established a pan-India research network and developed sophisticated tools to mine genetic data for exploratory research and drug target discovery
- Head and neck cancer case study identifies high frequently occurring mutations in Caspase8
- Case studies examining variants in pharmacogenes and SMN2 (important in Duchenne muscular dystrophy) show relevance of variants across multiple populations and reveal novel variants for further research, highlighting the value of the Indian genetic data