NEW COLOMBO PLAN

Connect to Australia's future - study in the region

Curtin University

Making Genetic Testing Reports Machine Readable

Haylee Jackson

Rare Care Centre

Clinical Centre of Expertise for Rare and Undiagnosed Diseases

SingHealth DukeNUS ACADEMIC MEDICAL CENTRE

Genomic Medicine Centre

SingHealth DukeNUS ACADEMIC MEDICAL CENTRE

Institute of Precision Medicine (PRISM)

KK Women's and Children's Hospital

SingHealth

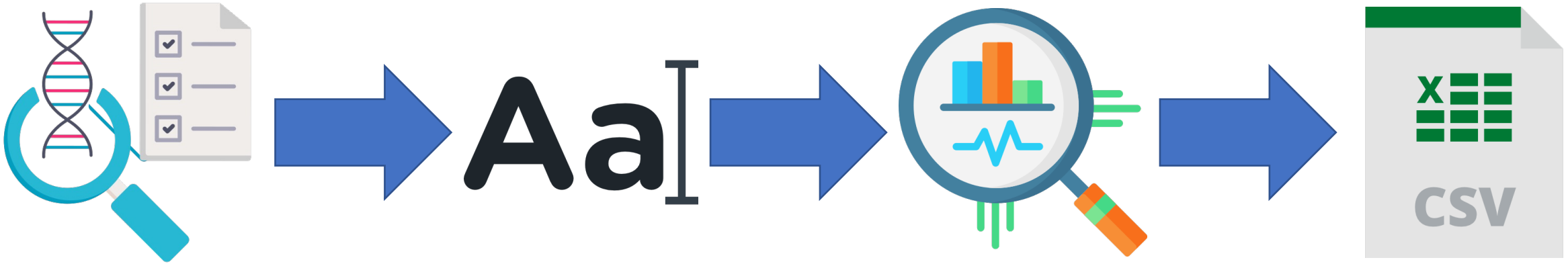
Problem

- The Genetic Testing Reports are PDFs
- Not integrated into Electronic Health Records (EHR)
  - A physician, who is not the ordering physician, will not have access to the genetic test report
  - Can lead to duplicate ordering in some instances
- Data needs to be manually extracted
- By fixing this we free up more time that can be better spent else where.
- A successful solution would reduce the amount of time spent on the extraction of the PDFs. It would also have the data in an easy way to filter and search.

Project Constraints & Assumptions

- 7 week project
- Choose to only extracted from Invitae Reports
  - Invitae is a lab used by both Australia and Singapore
  - KKH uses Invitae routinely for diagnostic genetic testing

The Plan



Genetic Testing Report PDF

Extracted Text

Key Data

Spreadsheet

PDF Extraction Tools Comparison

Tools	Extracts Text	Extracts Tables	Extracts Shape objects	Modify PDFs	Visual Debug Tools	Additional Installation Requirements
pdfplumber	✓	✓	✓	✓	✓	–
tika	✓	✗	✗	✓	✗	Requires Java 7+
pdfmef	✓	✓	✓	✗	✗	Requires Python 2.7
pdfminer	✓	✗	✓	✓	✗	–
PyPDF2	✓	✗	✗	✓	✗	–
pymupdf	✓	✗	✗	✓	✗	Requires MuPDF (non-Python software)
pdftables	✗	✓	✓	✗	✗	–

PDF Data & Patterns

- Using Regular Expressions (RegEx) to extract the data
- Extracted data included
  - Patient Details
  - Sample Type
  - Report Date
  - Report ID
  - Result
  - Gene Variation Details
  - List of Genes Tested

INVITAE

INVITAE DIAGNOSTIC TESTING RESULTS

Patient name:
DOB:
Sex assigned at birth:
Gender:

Sample type:
Sample collection date:
Sample accession date:
MRN:

Report date:
Invitae #:
Clinical team:

Reason for testing

Diagnostic test for a personal history of disease

Test performed

Sequence analysis and deletion/duplication testing of the 21 genes listed in the Genes Analyzed section.

- Invitae Overgrowth and Macrocephaly Syndromes Panel

RESULT: POSITIVE

One Pathogenic variant identified in PIK3R2. PIK3R2 is associated with autosomal dominant megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome.

Additional Variant(s) of Uncertain Significance identified.

GENE	VARIANT	ZYGOSITY	VARIANT CLASSIFICATION
PIK3R2	c.1117G>A (p.Gly373Arg)	heterozygous	PATHOGENIC
NSD1	c.2701A>G (p.Ile901Val)	heterozygous	Uncertain Significance

About this test

This diagnostic test evaluates 21 gene(s) for variants (genetic changes) that are associated with genetic disorders. Diagnostic genetic testing, when combined with family history and other medical results, may provide information to clarify individual risk, support a clinical diagnosis, and assist with the development of a personalized treatment and management strategy.

Clinical summary

A Pathogenic variant, c.1117G>A (p.Gly373Arg), was identified in PIK3R2.

- The PIK3R2 gene is associated with autosomal dominant megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome (MedGen UID: 861164).
- This result is consistent with a diagnosis of megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome.
- MPPH is a childhood-onset overgrowth syndrome characterized primarily by megalencephaly, hydrocephalus, and polymicrogyria. Additional features vary, and may include oromotor dysfunction, mild to severe intellectual disability, epilepsy, postaxial polydactyly, and connective tissue dysplasia (PMID: 22729224, 22228622).
- Biological relatives have a chance of being at risk for megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome and should consider testing if clinically appropriate.

A Variant of Uncertain Significance, c.2701A>G (p.Ile901Val), was identified in NSD1.

- The NSD1 gene is associated with autosomal dominant Sotos syndrome (MedGen UID: 833601).
- Not all variants present in a gene cause disease. The clinical significance of the variant(s) identified in this gene is uncertain. Until this uncertainty can be resolved, caution should be exercised before using this result to inform clinical management decisions.
- Familial VUS testing is not offered. Testing family members for this variant will not contribute evidence to allow variant reclassification. Details on our VUS Resolution and Family Variant Testing Programs can be found at https://www.invitae.com/family.

Variant details

PIK3R2, Exon 10, c.1117G>A (p.Gly373Arg), heterozygous, PATHOGENIC

- This sequence change replaces glycine, which is neutral and non-polar, with arginine, which is basic and polar, at codon 373 of the PIK3R2 protein (p.Gly373Arg).
- This variant is not present in population databases (gnomAD: no frequency).
- This missense change has been observed in individual(s) with megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome, or clinical features of this syndrome (PMID: 22729224, 24497998, 26520804, 28086757). In at least one individual the variant was observed to be de novo.
- ClinVar contains an entry for this variant (Variation ID: 39808).
- Advanced modeling of protein sequence and biophysical properties (such as structural, functional, and spatial information, amino acid conservation, physicochemical variation, residue mobility, and thermodynamic stability) performed at Invitae indicates that this missense variant is expected to disrupt PIK3R2 protein function.
- Experimental studies have shown that this missense change affects PIK3R2 function (PMID: 22729224).
- For these reasons, this variant has been classified as Pathogenic.

NSD1, Exon 5, c.2701A>G (p.Ile901Val), heterozygous, Uncertain Significance

- This sequence change replaces isoleucine, which is neutral and non-polar, with valine, which is neutral and non-polar, at codon 901 of the NSD1 protein (p.Ile901Val).
- This variant is present in population databases (rs753120157, gnomAD: 0.003%).
- This variant has not been reported in the literature in individuals affected with NSD1-related conditions.
- Advanced modeling of protein sequence and biophysical properties (such as structural, functional, and spatial information, amino acid conservation, physicochemical variation, residue mobility, and thermodynamic stability) performed at Invitae indicates that this missense variant is not expected to disrupt NSD1 protein function.
- In summary, the available evidence is currently insufficient to determine the role of this variant in disease. Therefore, it has been classified as a Variant of Uncertain Significance.

Genes analyzed

This table represents a complete list of genes analyzed for this individual, including the relevant gene transcript(s). If more than one transcript is listed for a single gene, variants were reported using the first transcript listed unless otherwise indicated in the report. Results are negative unless otherwise indicated in the report. Benign and Likely Benign variants are not included in this report but are available upon request. An asterisk (*) indicates that this gene has a limitation. Please see the Limitations section for details.

GENE	TRANSCRIPT
AKT2	NM_001626.5
AKT3	NM_005465.4
CDKN1C	NM_000076.2
CUL4B	NM_003588.3
DIS3L2*	NM_152383.4
DNMT3A	NM_175629.2
EZH2*	NM_004456.4
GLI3	NM_000168.5
GPC3*	NM_004484.3
KPTN	NM_007059.3
MED12	NM_005120.2
MTOR	NM_004958.3
NF1*	NM_000267.3
NFIX	NM_001271043.2
NPR2	NM_003995.3
NSD1	NM_022455.4
PHF6	NM_032458.2
PIK3R2	NM_005027.3
PTEN*	NM_000314.4
SETD2	NM_014159.6
SPRED1	NM_152594.2

INVITAE

INVITAE DIAGNOSTIC TESTING RESULTS

Patient name:
DOB:
Sex assigned at birth:
Gender:
Patient ID (MRN):

Sample type: Blood
Sample collection date:
Sample accession date:

Report date:
Invitae #:
Clinical team:

Reason for testing
Family history

Test performed
Sequence analysis and deletion/duplication testing of the 3 genes listed in the Genes Analyzed section.

REQUESTED VARIANTS

GENE	VARIANT	ZYGOSITY	VARIANT CLASSIFICATION	RESULT
HPS1	Deletion (Exons 7-17)	heterozygous	PATHOGENIC	Detected
AIRE	c.355C>A (p.Pro119Thr)	N/A	Uncertain Significance	Not detected
CARD14	c.1099G>T (p.Ala367Ser)	N/A	Uncertain Significance	Not detected
HPS1	c.425G>A (p.Arg142His)	N/A	Uncertain Significance	Not detected

The table above reflects the information for the requested variant(s) as of the date that this report was issued. Please see the result box for a summary of any reportable findings.

RESULT: CARRIER

One Pathogenic variant identified in HPS1. HPS1 is associated with autosomal recessive Hermansky-Pudlak syndrome 1.

GENE	VARIANT	ZYGOSITY	VARIANT CLASSIFICATION
HPS1	Deletion (Exons 7-17)	heterozygous	PATHOGENIC

About this test
This diagnostic test evaluates 3 gene(s) for variants (genetic changes) that are associated with genetic disorders. Diagnostic genetic testing, when combined with family history and other medical results, may provide information to clarify individual risk, support a clinical diagnosis, and assist with the development of a personalized treatment and management strategy.

Clinical summary

A Pathogenic variant, Deletion (Exons 7-17), was identified in HPS1.

- The HPS1 gene is associated with autosomal recessive Hermansky-Pudlak syndrome 1 (HPS1) (MedGen UID: 419514).
- This individual is a carrier for autosomal recessive HPS1. This result is insufficient to cause autosomal recessive HPS1; however, carrier status does impact reproductive risk.
- Hermansky-Pudlak syndrome is characterized by oculocutaneous albinism and bleeding diathesis caused by deficient dense granules in platelets (PMID: 9562579). Clinical manifestations include hypopigmentation of the skin and hair, reduced iris pigment, reduced retinal pigment, foveal hypoplasia, reduction in visual acuity, nystagmus, and increased crossing of the optic nerve fibers. Bleeding diathesis manifests as easy bruising, frequent epistaxis, colonic bleeding, menorrhagia, and prolonged bleeding after tooth extraction and other surgeries (PMID: 12125811). Solar keratoses and melanocytic nevi are common, and individuals may have an increased risk of squamous cell and basal cell carcinoma (PMID: 10411151). The HPS1 subtype is associated with severe oculocutaneous albinism and bleeding diathesis, lethal pulmonary fibrosis, and granulomatous colitis (PMID: 18544035).
- Biological relatives have a chance of being a carrier for or being at risk for autosomal recessive HPS1. Testing should be considered if clinically appropriate. The chance of having a child with autosomal recessive HPS1 depends on the carrier state of this individual's partner.

Variant details

HPS1, Deletion (Exons 7-17), heterozygous, PATHOGENIC

- This variant is a gross deletion of the genomic region encompassing exon(s) 7-17 of the HPS1 gene. This variant would be expected to be in-frame, preserving the integrity of the reading frame.
- This variant has not been reported in the literature in individuals affected with HPS1-related conditions.
- This variant disrupts a region of the HPS1 protein in which other variant(s) (p.Leu239Pro) have been determined to be pathogenic (PMID: 12442288, 29345414). This suggests that this is a clinically significant region of the protein, and that variants that disrupt it are likely to be disease-causing.
- For these reasons, this variant has been classified as Pathogenic.

INVITAE

INVITAE CARRIER SCREEN RESULTS

Patient name:
DOB:
Sex assigned at birth:
Gender:
Patient ID (MRN):

Sample type: gDNA
Sample collection date:
Sample accession date:

Report date:
Invitae #:
Clinical team:

Reason for testing

Patient/partner is planning for pregnancy

Test performed

Invitae Comprehensive Carrier Screen

- Primary Panel (CF, SMA)
- Add-on Comprehensive Carrier Screen genes

RESULT: POSITIVE

This carrier test evaluated 289 gene(s) for genetic changes (variants) that are associated with an increased risk of having a child with a genetic condition. Knowledge of carrier status for one of these conditions may provide information that can be used to assist with family planning and/or preparation. Carrier screening is not intended for diagnostic purposes. To identify a potential genetic basis for a condition in the individual being tested, diagnostic testing for the gene(s) of interest is recommended.

This test shows the presence of clinically significant genetic change(s) in this individual in the gene(s) indicated below. No other clinically significant changes were identified in the remaining genes evaluated with this test.

RESULTS	GENE	VARIANT(S)	INHERITANCE	PARTNER TESTING RECOMMENDED
Carrier: Polycystic kidney disease (PKHD1-related)	PKHD1	c.11740C>T (p.Arg3914*)	Autosomal recessive	Yes

Variant details

PKHD1, Exon 66, c.11740C>T (p.Arg3914*), heterozygous, PATHOGENIC

- This sequence change creates a premature translational stop signal (p.Arg3914*) in the PKHD1 gene. While this is not anticipated to result in nonsense mediated decay, it is expected to disrupt the last 161 amino acid(s) of the PKHD1 protein.
- This variant is present in population databases (rs761704401, gnomAD 0.02%).
- This variant has not been reported in the literature in individuals affected with PKHD1-related conditions.
- ClinVar contains an entry for this variant (Variation ID: 554930).
- This variant disrupts a region of the PKHD1 protein in which other variant(s) (p.Arg3961*) have been determined to be pathogenic (Invitae). This suggests that this is a clinically significant region of the protein, and that variants that disrupt it are likely to be disease-causing.
- For these reasons, this variant has been classified as Pathogenic.

Diagnostic Testing & Familial Variant Testing Results

Diagnostic Testing & Familial Variant Testing

Reason for Testing	Tests	Result	Gene	Transcript	Variant
Diagnostic test for a personal history of disease	Invitae Overgrowth and Macrocephaly Syndromes Panel	POSITIVE	PIK3R2	NM_005027.3	c.1117G>A
Diagnostic test for a personal history of disease	Invitae Overgrowth and Macrocephaly Syndromes Panel	POSITIVE	NSD1	NM_022455.4	c.2701A>G
Family history	Familial Variant Testing	CARRIER	HPS1	NM_000195.4	Deletion

Diagnostic Testing

Result	Gene	Transcript	Variant	Protein Change	Exon	Zygosity	Classification	Association	Population Data	Protein Modeling
POSITIVE	PIK3R2	NM_005027.3	c.1117G>A	p.Gly373Arg	10	heterozygous	PATHOGENIC	Autosomal do	no frequency	Expected to disrupt
POSITIVE	NSD1	NM_022455.4	c.2701A>G	p.Ile901Val	5	heterozygous	Uncertain Significance	Autosomal do	0.003%	Not expected to disrupt

Familial Variant Testing

Result	Gene	Transcript	Variant	Protein Change	Exon ▲	Zygosity	Classification
CARRIER	HPS1	NM_000195.4	Deletion	Exons 7-17	7-17	heterozygous	PATHOGENIC

Carrier Screening Results

Sample Type	Reason for Testing	Tests	Result	Gene	Transcript	Variant	Protein Change
gDNA	Patient/partner is planning for pregnancy	Primary Panel (CF SMA)	POSITIVE	PKHD1	NM_138694.3	c.11740C>T	p.Arg3914*

Exon	Zygosity	Classification	Carrier	Inheritance	Partner Testing
66	heterozygous	PATHOGENIC	Carrier:Polycystic kidney disease (PKHD1-related)	Autosomal recessive	Yes

Final Solution

```
~/Desktop/Singapore Project/Genetic-Report-
hayleej@MacBook-Air-3 Genetic-Report-PDF-Extraction % python3 main.p
```

Final Solution

Lab	Report Date	Sample Type	Reason for Testing	Tests	Result	Gene	Transcript	Variant	Protein Change
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Ketolysis Disorders Panel	NEGATIVE				
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Primary Immunodeficiency Panel	UNCERTAIN	CASP10	NM_032977.3	c.664A>G	p.Thr222Ala
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Primary Immunodeficiency Panel	UNCERTAIN	IKBKB	NM_001556.2	c.1849G>A	p.Val617Ile
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Primary Immunodeficiency Panel	UNCERTAIN	PSTPIP1	NM_003978.3	c.1183G>A	p.Gly395Ser
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Primary Immunodeficiency Panel	UNCERTAIN	TFRC	NM_003234.3	c.821T>C	p.Leu274Ser
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Disorders of Sex Development Panel	UNCERTAIN	CBX2	NM_005189.2	c.1006G>T	p.Val336Leu
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Aortopathy Comprehensive Panel	NEGATIVE				
Invitae		Blood	Family history	Familial Variant Testing	NEGATIVE				
Invitae		Blood	Family history	Familial Variant Testing	CARRIER	HPS1	NM_000195.4	Deletion	Exons 7-17
Invitae		Blood	Diagnostic test for a personal history of disease	Invitae Monogenic Diabetes Panel; 1 individual gene	UNCERTAIN	PPARG	NM_015869.4	c.1433T>G	p.Val478Gly
Invitae		gDNA	Diagnostic test for a personal history of disease	Invitae Connective Tissue Disorders Panel	NEGATIVE				
Invitae		gDNA	Family history	Familial Variant Testing	NEGATIVE				
Invitae		Blood	Family history	Familial Variant Testing	CARRIER	FANCA	NM_000135.2	c.987_990del	p.His330Alafs*4
Invitae		Blood	Family history	Familial Variant Testing	NEGATIVE				
Invitae		Blood	Family history	Familial Variant Testing	POSITIVE	LDLR	NM_000527.4	c.268G>A	p.Asp90Asn
Invitae			Diagnostic test for a personal history of disease	Invitae Overgrowth and Macrocephaly Syndromes Panel	POSITIVE	PIK3R2	NM_005027.3	c.1117G>A	p.Gly373Arg
Invitae			Diagnostic test for a personal history of disease	Invitae Overgrowth and Macrocephaly Syndromes Panel	POSITIVE	NSD1	NM_022455.4	c.2701A>G	p.Ile901Val

Conclusion

- Program takes on average 1.82 minutes for 18 reports
- Manual extraction takes on average 15 minutes per report
- Future Work
  - Other Genetic Testing Labs including Prevention Genetics
  - Working with Xiaohui and Sing Yi from Health Services Research Unit to establish if the script will run on the corporate network
  - Deploy program to extract the information for the identifiable reports
  - Integrate with EHR

Acknowledgements

- Dr Saumya Shekhar Jamuar
- Dr Sonny Pham
- Ms Sylvia Kam
- Ms Lim Jiin Ying
- Ms Jasmine Goh
- Dr Sonia Davila
- Dr Lim Weng Khong
- Ms Yasmin Bysltra
- Ms Simone Ng
- Ms Xin Xiaohui
- Ms Chia Sing Yi
- Ms Tang Yu Qun

CIRB 2019/2243

NEW COLOMBO PLAN

Connect to Australia's future - study in the region

Curtin University

Thank You!

Haylee Jackson

Rare Care Centre

Clinical Centre of Expertise for Rare and Undiagnosed Diseases

SingHealth DukeNUS ACADEMIC MEDICAL CENTRE

Genomic Medicine Centre

SingHealth DukeNUS ACADEMIC MEDICAL CENTRE

Institute of Precision Medicine (PRISM)

KK Women's and Children's Hospital

SingHealth