Streamlining Cytogenetic Data Processing with ISCN Parsing and OMOP

Ben Smith[1], Trent Peterson[1], Jessica Manzyuk[1]
[1] Principia Health Sciences, Cary, NC

Background

Karyotype data is important for cytogenetic research needed to improve diagnosis and treatment of numerous cancer types.[1] The International System for Human Cytogenomic Nomenclature (ISCN) is the central reference for the description of karyotyping, Fluorescence In Situ Hybridization (FISH), and microarray results, and provides rules for describing cytogenetic and molecular cytogenetic findings in laboratory reports.[2] Although a broadly adopted data standard, ISCN-formatted strings have been difficult to use in large-scale computational analyses.[3] Capturing cytogenetic results and integrating them with other clinical data commonly requires significant manual effort and is prone to human error in data entry and validation steps.

As an example, all US bone marrow transplant centers are required to submit cytogenetic information for every person receiving a hematopoietic stem cell transplant (HSCT) or cell therapy to a central registry maintained by the Center for International Blood and Marrow Transplant Research (CIBMTR). This collection activity can require multiple hours from chart abstractors who commonly lack specialized medical training. The resulting data also applies to limited research use cases because of incompleteness (i.e., CIBMTR doesn't require that all abnormalities be identified and submitted).

Our team explored ways to streamline and automate steps of this process while producing data that could easily be integrated with other clinical information in support of large-scale studies. The resulting approach involved automatic parsing of ISCN strings and storing the resulting values in Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

Methods

To support automation of data extraction from ISCN strings, we developed an open-source ISCN string parser using the ANTLR (ANother Tool for Language Recognition) programming language. Our tool isn't the only one built for this purpose, but we encountered limitations in other parsers that motivated us to build a new one with greater flexibility and comprehensiveness. The parser identifies specific abnormalities and other information contained in the ISCN strings using ANTLR-based decision tree logic. It then stores the results in JSON format for additional processing. We selected ANTLR to develop the parsing logic because it aligns very well with the complex structure of ISCN strings (e.g., it can handle greater complexity than regular expressions are capable of). In preparation for research analysis, the JSON-based ISCN strings and parsed cytogenetic abnormality values are then converted and stored in OMOP CDM.

During development of this capability, it was not clear where in OMOP CDM the values should best be stored. Most fields in OMOP cannot store more than 50 characters, which would preclude storage of longer ISCN strings. After failing to find evidence of validated prior art and consulting with experts, we determined that NOTE (for the strings) and NOTE_NLP (for the parsed values) were best suited due to absence of character count limits in NOTE and the data type alignment with NOTE_NLP expectations from researchers.

To test accurate string parsing and storage of values, the team leveraged a series of complex ISCN strings developed by clinical experts and manually audited the resulting data and where it was stored.

Results

Our comparison of source ISCN strings and processed results confirmed that the parser successfully recognized all target abnormalities within complex strings and stored them in appropriate OMOP CDM tables—both the string (in NOTE) and the parsed abnormality values (in NOTE_NLP). This data could be analyzed alongside corresponding clinical information for given synthetic patient records.

Conclusion

This automated approach to cytogenetic data processing represents a significant opportunity to scale cytogenetic research by reducing manual effort required for data entry and processing. Future use of this streamlined approach could enable inclusion of cytogenetics data in a wider range of studies.

Ideally, researchers from the OHDSI community would be able to use ATLAS, OHDSI's tool for descriptive analyses, with this parsed data. Currently, ATLAS does not work properly with NOTE and NOTE_NLP tables. We are currently scoping development of this capability and expect to contribute updates to ATLAS for use by the OHDSI community.

We are also experimenting with the use of optical character recognition and natural language processing for extracting ISCN strings from PDF and printed lab reports. Early exploration suggests great promise for further automating this process using these technologies.

References

1. McGowan-Jordan J. et al. (2016) ISCN 2016: An International System for Human Cytogenomic Nomenclature. Karger Medical and Scientific Publishers, Basel.

2. Stevens-Kroef M, Simons A, Rack K, Hastings RJ. Cytogenetic Nomenclature and Reporting. Methods in Molecular Biology. 2016 Dec 2;303–9.

3. Abrams ZB, Tally DG, Abruzzo LV, Coombes KR. RCytoGPS: An R package for reading and visualizing cytogenetics data, Bioinformatics. 2021;37(23): 4589–4590.

Figure 1 Example ISCN String Parsing Analysis

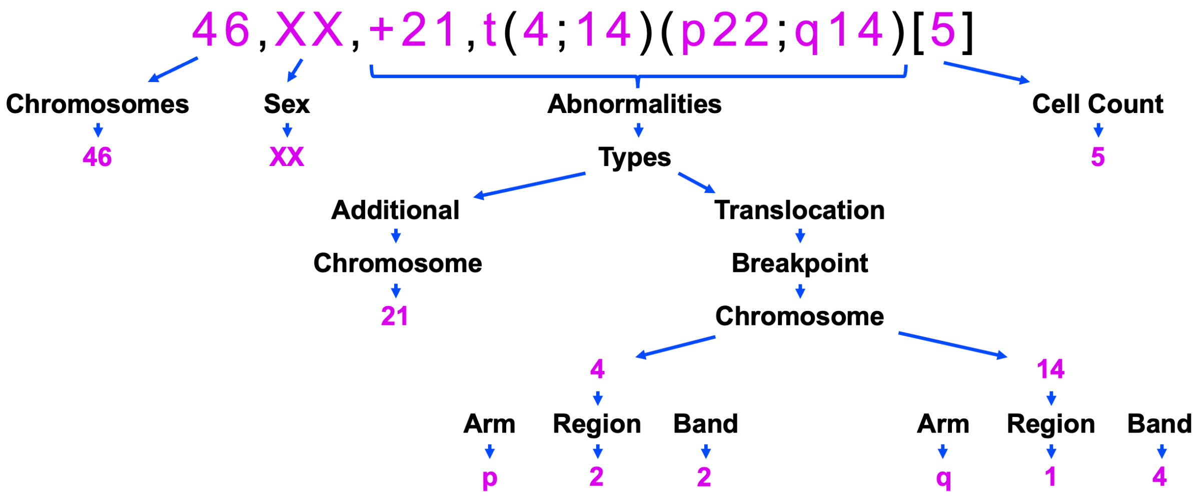
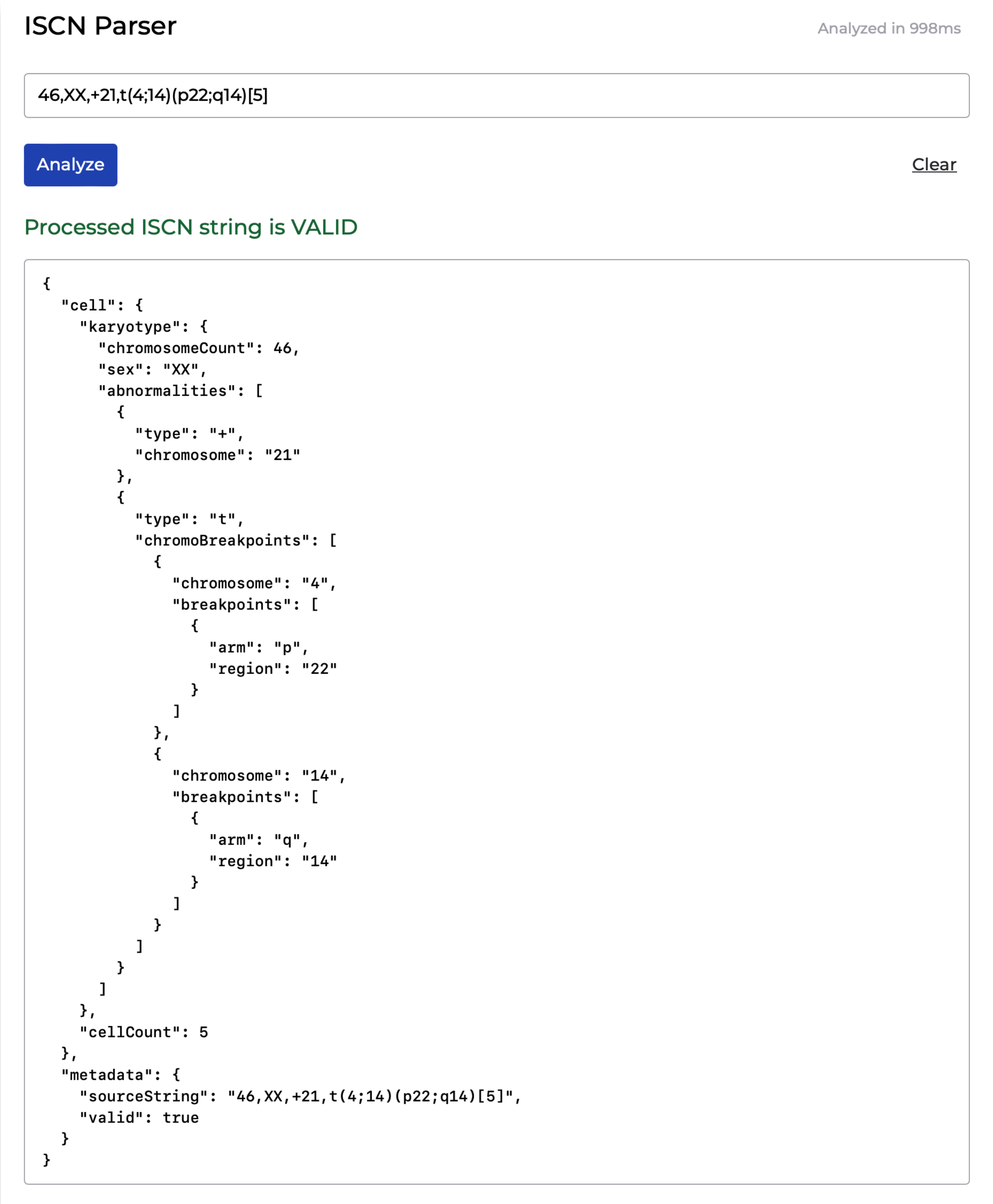


Figure 2 Example Parsed String Validation and JSON output



ISCN Parser

Analyzed in 998ms

46,XX,+21,t(4;14)(p22;q14)[5]

Processed ISCN string is VALID

```
{
  "cell": {
    "karyotype": {
      "chromosomeCount": 46,
      "sex": "XX",
      "abnormalities": [
        {
          "type": "+",
          "chromosome": "21"
        },
        {
          "type": "t",
          "chromoBreakpoints": [
            {
              "chromosome": "4",
              "breakpoints": [
                {
                  "arm": "p",
                  "region": "22"
                }
              ]
            },
            {
              "chromosome": "14",
              "breakpoints": [
                {
                  "arm": "q",
                  "region": "14"
                }
              ]
            }
          ]
        }
      ]
    },
    "cellCount": 5
  },
  "metadata": {
    "sourceString": "46,XX,+21,t(4;14)(p22;q14)[5]",
    "valid": true
  }
}
```