Aster Insights™

www.AsterInsights.com

APRIL, 2023

WHITE PAPER

CONSENSUS MOLECULAR SUBTYPE ANALYSES IN COLORECTAL CANCER SAMPLES

Authors:
David M. McKean, PhD
Daniel Elgort, PhD
Oliver A. Hampton, PhD, MS
Phaedra Agius, PhD

ABSTRACT

Consensus molecular subtyping (CMS) of colorectal cancer (CRC) samples has increased understanding of the established subtypes on both a phenotypic and genotypic level and has enabled predicted outcomes for more personalized treatments. As a validation of both the quality, breadth and depth of Aster Insights' real world clinical and molecular data, CMS groups were predicted for CRC tumor samples and then evaluated and compared with known features of the CRC consensus molecular subtypes.

Introduction

Preventative screening for colorectal cancer (CRC), resulting in early detection and surgical removal prior to metastasis, has dramatically decreased the CRC mortality rate. However, metastatic CRC remains the second leading cause of cancer-related mortality. Personalized treatments of CRC with targeted therapies offer reduced mortality, but they are currently limited to a handful of prognostic markers. Anti-epidermal growth factor receptor (EGFR) monoclonal antibodies (mAbs) are effective in treating cancers with wild-type KRAS. Immune checkpoint inhibitors (anti-programmed cell death protein 1 (PD-1) mAbs) have proven effective against cancers with microsatellite instability. New and more effective personalized treatments will likely require a deeper understanding of CRC subtypes.

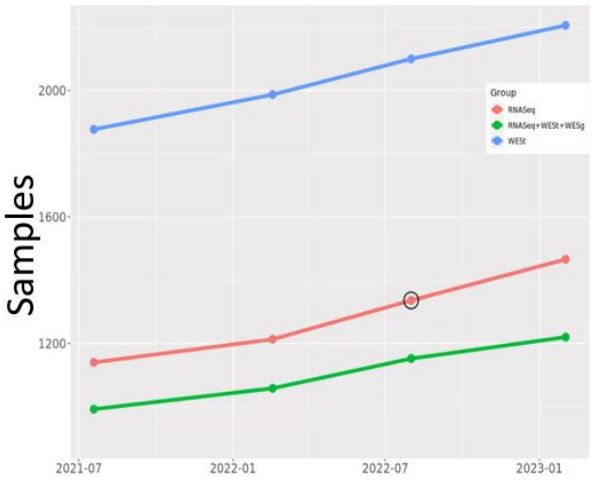
The CRC community first defined consensus molecular subtypes (CMS) in a groundbreaking publication by Guinney, et al. in 2015*[1]*. The authors described four consensus subtypes consisting of (1) MSI Immune, (2) Canonical, (3) Metabolic and (4) Mesenchymal based on analysis of 4,151 colorectal cancers from 18 different data sets. Characterization of these subtypes at both the biological level*[2]* and retrospectively, at the treatment level*[3,4]*, have demonstrated CMS classification as a useful tool in personalized treatment. Currently, one of the obstacles prohibiting the use of CMS for targeted therapy is the requirement for clinical RNASeq and downstream analysis.

Aster Insights is a patient-first, oncology-focused health informatics solutions company that seeks to accelerate discovery, development and delivery of more personalized therapies to improve patient outcomes. Aster Insights is well-positioned to support both academic research (by driving scientific collaborations and expanding cancer knowledge) *[5,6]* and pharmaceutical research, by expediting time-to-market drug discovery*[7]*. Aster Insights generates molecular data (whole exome sequencing and RNAseq) and integrates it with real-world, longitudinal clinical data in order to improve cancer care by accelerating precision medicine. Herein, we describe the subtyping and characterization of the colorectal cancer samples that exist in the current Aster Insights database as a demonstration of the quality, breadth, and depth of Aster Insights clinical and molecular data. This work describes known features of the CRC consensus molecular subtypes and is intended to showcase Aster Insights data.

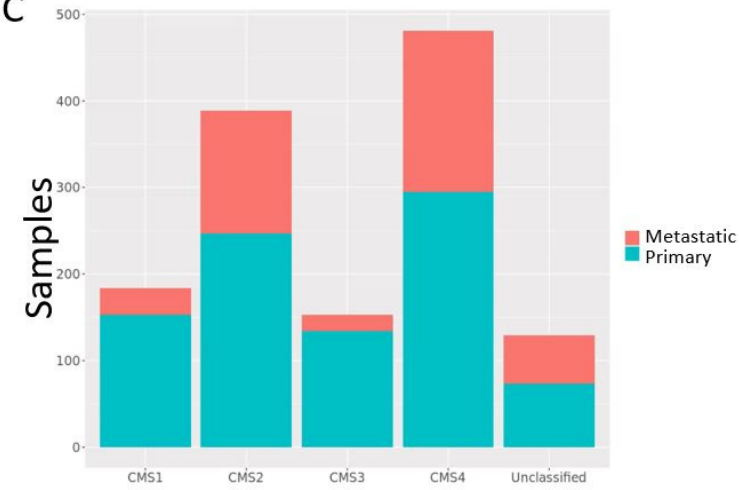
Results

To evaluate Aster Insights' real world clinical and molecular data, CRC samples were assigned to one of four CMS groups. 1336 primary or metastatic CRC samples (1228 Avatars; see Methods for Avatar description) with RNASeq expression data (from Aster Insights' October 2022 release; Fig. 1A) were assigned to CMS groups using the R package "CMScaller", which makes assignments based on the expression of 473 informative, pre-selected genes*[8]*. 1207 samples (1108 Avatars) were successfully assigned to the four well-defined subtypes, while 129 samples (120 Avatars) remained "unclassified" (Fig. 1B). CMS4 samples were the most abundant (n=481), followed by CMS2 (n=389) then CMS1 (n=184) and CMS3 (n=153). Due to the inclusion of metastatic samples which are highly represented in CMS4 (Fig. 1C), and the use of a different CMS classifier, CMS4 was over-represented (36% vs. 23%) in Aster Insights data, as compared to Guinney, et al*[1]*.

A



C



B

	Aster Insights Samples		Aster Insights Avatars		Guinney
	n	Pct	n	Pct	Pct
CMS1	184	14%	178	14%	14%
CMS2	389	29%	365	30%	37%
CMS3	153	11%	146	12%	13%
CMS4	481	36%	419	34%	23%
Unclassified	129	10%	120	10%	13%
Total	1336		1228		

Figure 1:
Molecular subtyping of Aster Insights colorectal cancer samples.
(A) Shown are the number of Aster Insights' colorectal cancer samples over time. The number of samples with tumor whole exome sequencing (WESt; blue), tumor RNASeq (RNASeq; red) or tumor and normal WES and tumor RNASeq (RNASeq+WESt+WESg; green) are indicated. The black circle highlights 1336 samples with RNASeq that were used for these analyses. (B) Consensus molecular subtypes (CMS1/2/3/4 or Unclassified) were assigned using the R package CMScaller, and both number (n) and percentage (Pct) of samples are reported. Subtypes CMS1/2/3 were observed in similar percentages as in Guinney, et al. (2015); Subtype CMS4 was elevated. Shown in (C) are the number of primary cancers (blue) and metastatic cancers (red) by CMS group. Of note, the highest percentage of metastatic cancers occur in CMS4.

Results

After assigning CRC samples to CMS groups, we further characterized CRC samples with available tumor whole exome sequencing (WESt) data for microsatellite instability (MSI) and for somatic pathogenic mutations in CRC-related genes, such as BRAF and KRAS. Samples with germline WES (WESg) were additionally assessed for somatic tumor mutation burden (TMB) and pathogenic driver mutations among the mismatch repair genes. These data (MSI/TMB table, and ClinVar-annotated vcf files; see Methods) were all available as standard output of the Aster Insights' molecular pipeline. Longitudinal clinical data was used to assess gender, age, tumor location, etc. To assess overall survival, we restricted analyses to a single sample per Avatar. The sample size was thus reduced to 939 CMS1/2/3/4 samples (Table 1).

CMS1 (aka MSI Immune subtype) samples are the most well-characterized of the four CRC subtypes; hallmark features include MSI and high TMB*[1.]* As expected, Aster Insights' CMS1 samples are enriched for both MSI and high TMB (Table 1, Fig. 2A). We validated the known enrichment of pathogenic BRAF mutations in CMS1 (Table 1)*[1]*. Further, CMS1 samples are enriched in females (66%), are often derived from right-sided lesions (74%) and are more frequently at a reduced stage, as compared to CMS2/3/4 (Table 1, Fig. 2A). Owing to high TMB and increased efficacy, CMS1 samples are more likely to be treated with Immune Checkpoint Inhibitors (Table 1)*[9]*.

	CMS1	CMS2	CMS3	CMS4
Avatars (n)	149	315	121	354
Diagnosis Age				
Median (yrs)	64.5	59.2	63.2	52.8
Mean (yrs)	62.9	59.7	60.3	54.3
Under 45 (%)	10.1	7.9	10.7	20.1
Gender				
Female (%)	66.4	43.5	48.8	44.4
Stage				
I (%)	9.4	7.6	9.9	5.6
II (%)	30.9	22.9	32.2	18.4
III (%)	44.3	33.7	34.7	39.5
IV (%)	14.8	35.2	23.1	36.4
Location				
Left–sided (%)	26.4	80.1	52.2	72.8
Right–sided (%)	73.6	19.9	47.8	27.2
MSI Status				
MSI (%)	67.2	0.4	12.1	3.1
MSS (%)	32.8	99.6	87.9	96.9
PathMutations				
KRAS mut (%)	26.2	27	57	32.5
BRAF mut (%)	36.9	1.6	5	2.3
ICI Treatment				
Treated (%)	6	1.9	3.3	3.1

Table 1 Demographics

Results

CMS2 (aka Canonical subtype) samples are slightly enriched in males (57%), are often derived from left-sided lesions (80%) and are more frequently detected at late stage (Table 1). Aster Insights provides pre-calculated Gene Set Variation Analysis signatures (see Methods), so we assessed signatures known to be significantly associated with CMS2. As expected, the BIOCARTA Cell Cycle gene signature is significantly enriched in CMS2 (Fig. 2A)*[1]*. CMS2 has previously been associated with high somatic copy number alteration (SMCA)*[1]*; we validated this by plotting mean somatic copy number (of CMS groups) for each gene, observing consistently higher copy number of CMS2 samples (Fig. 2B).

CMS3 (aka Metabolic subtype) is the only subtype that is equally derived from left- and right-sided lesions and has no gender bias. CMS3 is enriched for MSI (compared to CMS2/4; Table 1). Further, CMS3 samples are enriched for pathogenic KRAS mutations (Table 1)*[1]* and have significantly enriched gene signatures for metabolic pathways (KEGG Fatty Acid Metabolism; Fig. 2A).

(Figure on next page)

Results

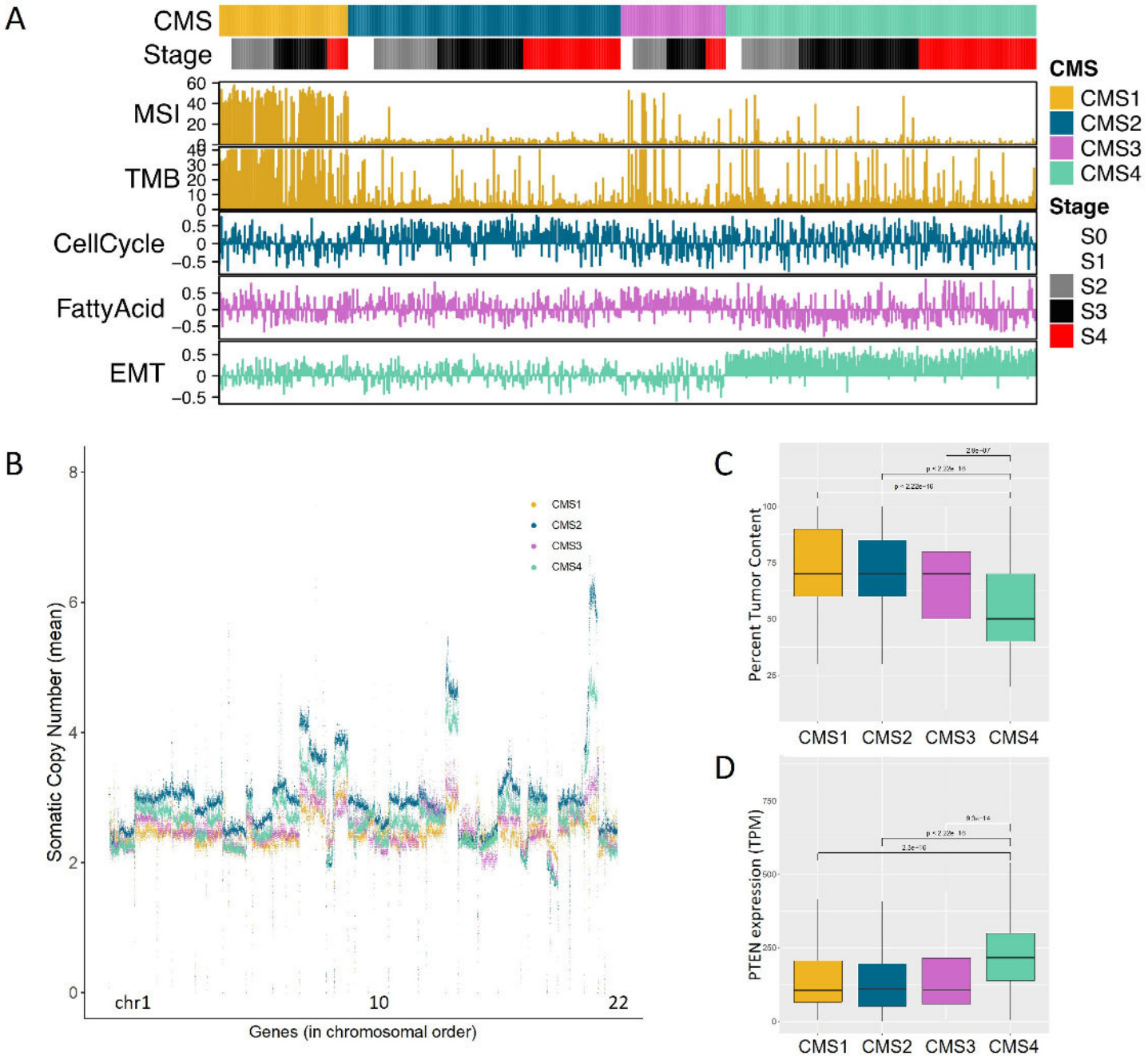


Figure 2:
Characterization of Aster Insights' CMS groups.
(A) ComplexHeatmap display of CRC samples, ordered on the x-axis by CMS (CMS1: yellow, CMS2: blue, CMS3: pink, CMS4: green), then by stage. MSI scores and TMB (yellow) are significantly enriched in CMS1, and are colored accordingly. The CellCycle (blue), FattyAcid (pink), and EMT (green) gene signatures are significant in CMS2, CMS3 and CMS4, respectively. Somatic copy number alteration is graphed in (B), with the mean somatic copy number (per CMS group) plotted on the y-axis and 57,438 genes on chromosomes 1-22 (in order) plotted on the x-axis. Percent tumor content by CMS is displayed by box and whisker plots (C), and the significantly reduced percent tumor content in CMS4 (p-values indicated) is a proxy for its invasiveness. PTEN expression by CMS is displayed by box and whisker plots (D), and CMS4 PTEN expression is significantly upregulated (p-values indicated).

Results

CMS4 (aka Mesenchymal subtype) samples are slightly enriched in males (56%) and are often derived from left-sided lesions (73%; Table 1). As compared to CMS1/2/3, CMS4 samples are more frequently associated with late stage (76% at Stage III/IV) and are diagnosed at an earlier age (20% diagnosed before 45; Table 1). CMS4 samples are reported to be more invasive than other CRC subtypes; we confirm this in Aster Insights' data as significantly reduced percent tumor content -a proxy for increased invasiveness (p-value < 2.9e-7; Fig. 2C). Loss of expression of the tumor suppressor gene PTEN is associated with over-activation of the PI-3 kinase/AKT pathway resulting in increased tumorigenesis and invasiveness*[10]*. We evaluated the RNASeq expression of PTEN among subtypes and found that PTEN is significantly upregulated in CMS4 (p-value < 8.5e-8; Fig. 2D).

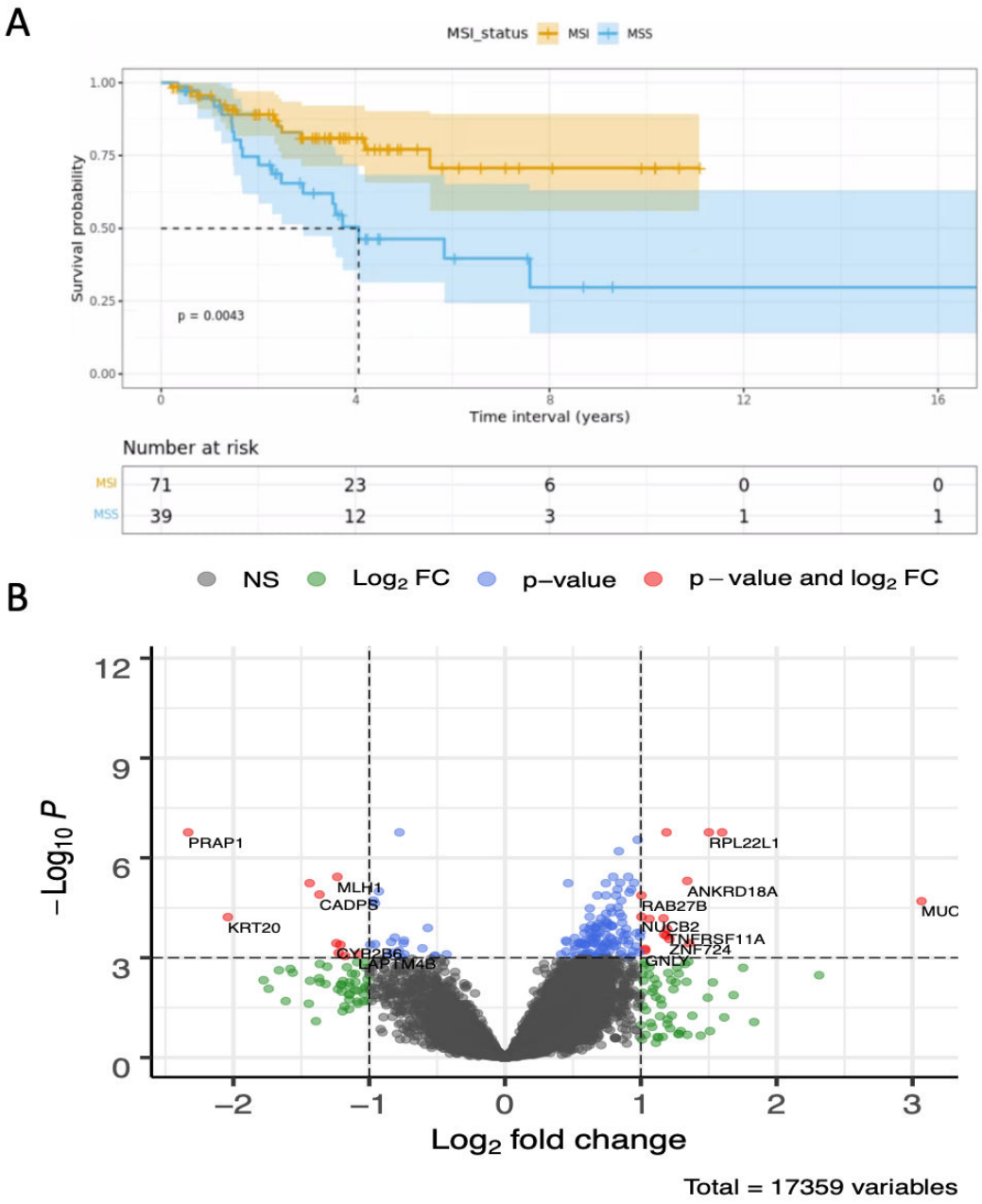


Figure 3: Overall survival and expression differences in CMS1 dependent on microsatellite instability. (A) Kaplan-Meier curve differentiating CMS1/microsatellite-unstable (blue) and CMS1/microsatellite-stable (yellow) samples reveal significantly better survival in CMS1 samples with microsatellite instability. (B) EnhancedVolcano plot, showing limma-based differential expression between CMS1/microsatellite-unstable and CMS1/microsatellite-stable samples. Genes with significant differential expression (p-value < 0.001) are represented as substantial fold change (|log2 fold change| > 1; red dots) or minimal fold change (blue). Genes that are not differentially expressed are in black or green.

Results

CMS1 samples were then ascertained for differences in overall survival based on MSI status. CMS1/MSS (microsatellite stable) samples (n=39) versus CMS1/MSI samples (n=71) were plotted on a Kaplan-Meier curve, which showed that median CMS1/MSS survival was ~4 years, at which time the CMS1/MSI samples had >80% survival rate (p-value=0.0043). To shed some light on the mechanism behind the significant difference in overall survival, we next evaluated gene expression differences using limma*[11]*. Nine and fourteen genes were significantly down- and up-regulated, respectively (|log2fold| > 1; p-value < 0.001). Included among these genes is the mismatch repair gene MLH1, a central player in microsatellite instability.

Conclusions

The Aster Insights clinical data model comprises of the following clinical data, as tables: (1) Cytogenetic abnormalities, (2) Diagnosis, (3) Family history, (4) Imaging, (5) Labs, (6) Medications, (7) Metastatic disease, (8) Outcomes, (9) Patient history, (10) Patient master, (11) Physical assessment, (12) Radiation, (13) Stem cell transplant, (14) Surgery biopsy, (15) Tumor marker, (16) Tumor marker flow panel, (17) Tumor sequencing, (18) Vital status and (19) Clinical Molecular Linkage file.

The Aster Insights molecular data model comprises of the following: (1) Tumor whole exome sequencing (bam and vcf) files, (2) Normal whole exome sequencing (bam and vcf) files, (3) RNASeq expression (gene and transcript) files, (4) QC and metrics files, (5) MSI/TMB file, (6) CNV file, (7) Gene fusion file (RNASeq-derived). Aster Insights will soon provide (8) structural variants.

Aster Insights offers two strategic partnerships with the pharmaceutical industry: (1) data licensing, in which industry partners receive both clinical and molecular data or (2) SEARCH, in which Aster Insights' bioinformatics team partners with individual industry partners to answer specific inquiries and investigate proposed hypothesis.

Aster Insights enables the greatest opportunity for oncology discovery researchers by using whole exome sequencing (DNA), whole transcriptome sequencing (RNA) and germline DNA sequencing of patient samples as its baseline. The company built and manages the Oncology Research Information Exchange Network (ORIEN®), a federated consortium of leading U.S. cancer centers, to promote increased collaboration in research and clinical trials. Most importantly, Aster Insights' patient-centric structure is based on lifetime patient consent using the Total Cancer Care® (TCC) protocol, making this one of the first longitudinal cancer patient databases of its kind.

Contact our Business Development team at Sales@AsterInsights.com for more information about our unique clinical and molecular dataset, and discover how we can help enable new insights for researchers and accelerate target identification and drug discovery using advanced patient cohorts.

Learn more at www.AsterInsights.com, and follow us on LinkedIn and Twitter.

Methods

ORIEN AVATAR Program

ORIEN AVATAR® is fully annotated (and continually updated) longitudinal clinical data, supported by whole exome and RNA sequencing data. A single Avatar may have multiple molecular samples (primary, metastatic, different time points), allowing unique study opportunities. The Aster Insights' database currently has ~22,000 Avatars.

Whole Exome Sequencing

Solid tumor samples are assessed by a clinical pathologist and are required to be > 30% tumor. DNA is purified from frozen or paraffin embedded tissue and normal DNA from blood or buccal swab (normal) using Qiagen QIASymphony. Libraries are generated, then captured on an Aster Insights-designed capture array prior to being sequenced on Illumina NovaSeq 6000. Mean target coverage is 300X for tumor samples and 100X for samples. WES data is aligned to GRCh38/hg38 genome using Sention, and output files (cram, vcf, metrics) are provided to Aster Insights' partners.

Downstream DNA Analysis

Microsatellite instability is scored using MSIsensor*[12]*. Tumor mutation burden is calculated as the sum of PASSing somatic (missense, nonsense and frameshift) variants divided by the exome capture region (in Mb). Copy number variation is assessed by Sequenza*[13]*. These analyses are part of the Aster Insights molecular data pipeline and are provided to Aster Insights' partners.

RNA Sequencing

Solid tumor samples are assessed by a clinical pathologist and are required to be > 30% tumor. RNA is purified from frozen tissue using QIAgen RNeasy plus mini kit, or from formalin-fixed paraffin-embedded tissue using the Covaris Ultrasonication FFPE DNA/RNA kit. RNASeq libraries are generated using Illumina TruSeq RNA Exome with single library hybridization and sequenced on Illumina NovaSeq 6000 to a coverage of ~100 million paired end reads. RNASeq data is aligned to GRCh38/hg38 using STAR, expression analysis is performed with RSEM and output files (cram, expression at both gene- and transcript-level counts, TPM, and FPKM) are provided to Aster Insights' partners.

Downstream RNA Analytics

Aberrent RNA fusions are assessed using STAR-fusion and Arriba*[14,15]*. Gene Set Variation Analysis (GSVA) is calculated for more than 11,000 gene signatures. These analyses are part of the Aster Insights molecular data pipeline and are provided to Aster Insights' partners.

Differential Expression

Minimally expressed genes (mean > 0.5 TPM; max > 1 TPM) were analyzed for differential expression using limma*[11]*. Results are displayed using the R package "EnhancedVolcano".

References

1. Guinney, et al. The Consensus Molecular Subtypes of Colorectal Cancer. Nat Med. (2015). PMID: 26457759.

2. Sawayama, et al. Investigation of colorectal cancer in accordance with consensus molecular subtype classification. Ann Gastroenterol Surg. (2020). PMID: 33005848.

3. Stintzing, et al. Consensus molecular subgroups (CMS) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. Annals of Oncology (2019). PMID: 31868905.

4. Lenz, et al. Impact of Consensus Molecular Subtype on Survival in Patients with Metastatic Colorectal Cancer: Results from CALGB/SWOG 80405 (Alliance). J Clin Oncol. (2019). PMID: 31042420.

5. Morris, et al. Replicative Instability Drives Cancer Progression. Biomolecules (2022). PMID: 36358918.

6. Hoskins, et al. Pan-cancer Landscape of Programmed Death Ligand-1 and Programmed Death Ligand-2 Structural Variations. JCO Precis Oncol. (2023). PMID: 36623238.

7. Ayers, et al. Molecular Profiling of Cohorts of Tumor Samples to Guide Clinical Development of Pembrolizumab as Monotherpy. Clin Cancer Res. (2019). PMID: 30442684.

8. Eide, et al. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. (2017). PMID: 29192179.

9. Andre, et al. Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer. New Engl J Med. (2020). PMID: 33264544.

10. Haddadi, et al. PTEN/PTENP1: 'Regulating the regulator of RTK-dependent PI3K/Akt signaling', new targets for cancer therapy. Mol Cancer. (2018). PMID: 29455665.

11. Ritchie, et al. Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. Nucleic Acids Res. (2015). PMID: 25605792.

12. Niu, et al. MSIsensor: Microsatellite Instability Detection Using Paired Tumor-Normal Sequence Data. Bioinformatics (2014). PMID: 24371154.

13. Favero, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. (2015). PMID: 25319062.

14. Uhrig et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. (2021). PMID: 33441414.

15. Dobin, et al. STAR: Ultrafast Universal RNA-seq Aligner. Bioinformatics (2018). PMID: 23104886.