Genomic landscape of atypical adenomatous hyperplasia reveals divergent modes to lung adenocarcinoma

Smruthy Sivakumar, F. Anthony San Lucas, Tina L. McDowell, Wenhua Lang, Li Xu, Junya Fujimoto, Jianjun Zhang, P. Andrew Futreal, Junya Fukuoka, Yasushi Yatabe, Steven M. Dubinett, Avrum E. Spira, Jerry Fowler, Ernest T. Hawk, Ignacio I. Wistuba, Paul Scheet*, Humam Kadara*

*Equally contributing co-corresponding authors

SUPPLEMENTARY METHODS

Specimen collection and evaluation: Specimens were obtained formalin-fixed and paraffin-embedded (FFPE) and stained (by hematoxylin and eosin/H&E). Assessment of histopathology of AAHs and LUADs was performed by analysis of H&E stained alternating slides (from 5 micron sections) with sections in between (10 micron) preserved for RNA isolation. Tissues were pathologically examined following the World Health Organization on the classification of lung tumors in the report by Travis and colleagues [1]. Images of the lesions were scanned using the Aperio platform (Leica Biosystems).

DNA and RNA isolation: For DNA and RNA isolation, 5 to 15 sections/slides per specimen were deparaffinized prior to isolation of normal tissues and lesions by scraping with 25-gauge needles under a stereomicroscope. Tissue fragments were then collected in 1.5 ml self-lock tubes containing 100 to 200 µl of lysis buffer PKD (Qiagen).

Deep targeted DNA sequencing: Target amplification was carried out in 5 µl reactions with 17 cycles of amplification. The pools were then combined for digestion and ligation using Ion Xpress barcode adapters (Thermo Fisher Scientific) according to the manufacturer's protocol. The libraries were then quantitated with quantitative PCR (qPCR) using the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific). Template reactions were prepared using the Ion PI Hi-Q OT2 200 Kit and the Ion PI Hi-Q Chef kit (Thermo Fisher Scientific) based on the commercial protocol. Templates were then assessed on a Qubit 2.0 fluorometer before loading onto Ion PI chip v3 (Thermo Fisher Scientific). For sequencing, specimens from two cases were processed together in one chip and sequenced on an Ion Proton sequencer. Sequencing reports generated in the Ion Torrent Suite 5.0 were used to assess the quality of the libraries and sequencing runs. Base calling results were aligned to the reference *hg19.fasta* provided by the manufacturer. The aligned BAM files were then used to run Torrent Variant Caller 5.0 (TVC) using manufacturer's targeted region BED file to generate VCF files.

Identification of somatic mutations: First, automated analysis was performed with Ion Torrent proprietary software Ion Reporter to call somatic variants in LUADs and AAHs (Table S1) by contrasting

events in NLs. The two available tumor specimens from case 20 (Table S1), were pooled together for analysis. The following two programs were used to augment variant calls in AAHs and LUADs: MuTect [2], using the default settings with the exception of retaining those annotated with the tag 'nearby_gap_events'; and Varscan2 [3], using a minimum variant allele frequency (VAF) threshold of 0.01987 instead of its default of 0.2 and filtering variants with a *P*-value < 10^{-6}. The VAF threshold for Varscan2 was reduced to identify a larger number of low frequency mutations given the high average depth of sequencing; the threshold of 0.01987 was derived based on the minimum VAF detected in Ion Reporter, the caller natively calibrated for this platform. Finally as a fourth caller, the VCF files generated marginally from Torrent Variant caller (TVC) were subjected to a simple subtraction of variants, removing those observed in the matched normal samples for each case. Potential homopolymers were filtered out using a custom script (http://scheet.org/) that identifies homopolymers in the vicinity of the mutation based on its genomic location. A default minimum homopolymer length of 6bp and a window size of 10bp on either side of the mutation were used. Short stretches of homopolymers involving the mutation identified were also examined. Variants from all four callers/methods were annotated using ANNOVAR [4] and a post-processing protocol was then applied to remove potentially-overlooked germline variants based on their presence in the 1000 genomes project [5] and the Exome Variant Server [6].

Validation of somatic mutations using digital PCR: Somatic mutations identified in the driver genes *BRAF* and *KRAS* in AAHs as well as the samples (AAH and LUAD) carrying the *EGFR* p.L858R mutation as detected in DNA targeted sequencing were verified by digital PCR using the QuantStudio 3D system (Thermofisher, A26317) following the manufacturer's protocol for use with a chip loader. TaqMan and Custom TaqMan SNP genotyping assays were used as probes for the mutations (Thermofisher 4351379 and 4332077). Samples with less than 30 ng of input DNA were given 7 cycles of preamplification using the Platinum™ PCR SuperMix High Fidelity (Thermofisher 12532016) and SNP genotyping assays as primers. Allele frequencies were obtained from analyzing the chip files in the QuantStudio 3D software available on the Thermofisher cloud.

Transcriptome sequencing: For library preparations, approximately 30 ng of RNA was used based on sample concentrations obtained from the Qubit HS RNA assay (Thermo Fisher Scientific). All samples were heat shocked at 80°C for 10 minutes and cooled to room temperature for 5 minutes and then reverse-transcribed to generate cDNA libraries using the Ion AmpliSeq Transcriptome Human Gene Expression Kit (Thermo Fisher Scientific) adhering to the manufacturer's protocol for FFPE samples with 16 cycles of target amplification. Library concentrations were determined by qPCR using an absolute quantitation method and the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific) following the manufacturer's protocol. Template reactions were carried out using the Ion PI Hi-Q OT2 200 Kit (Thermo Fisher Scientific) according to the manufacturer's instructions and then loaded onto Ion PI chips v3 using the Ion PI Hi-Q Sequencing 200 Kit based on the manufacturer's protocol (Thermo Fisher Scientific). The Ion Torrent Suite 5.0 was used to assess the quality of the libraries and sequencing runs. Results from base calling results were aligned to a reference file (*hg19_ampliseq_transcriptome_ercc_v1.fasta*), which produced aligned BAM files in the Ion Torrent server.

Expression analysis: Differentially expressed genes were identified by ANOVA based on a $P < 0.001$, false-discovery rate (FDR) of 0.01 and minimum of 2-fold change in at least one of three comparisons (AAH-NL, LUAD-NL and LUAD-AAH). The model incorporated specimen type as a fixed effect and with different patients considered random effects. Disparate patterns of differential expression from normal to AAH to LUAD were determined based on a trivial classifier composed of two one-sided *t*-tests ($P < 0.05$). To understand immune signaling in the pathogenesis of AAH, the expression of an *a priori* list of markers of immune response and function (nCounter PanCancer Immune Profiling Panel from nanoString technologies), was similarly analyzed but with a minimum 1.5-fold change required in at least one of the three comparisons mentioned above. For identifying genes differentially expressed among different groups of AAHs based on select mutation status, a $P < 0.01$ threshold and a 1.5-fold change cut-off was used.

REFERENCES

1. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol. 2015;10: 1243–1260.

2. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31: 213–219.

3. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22: 568–576.

4. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164–e164.

5. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. Nature Research; 2015;526: 68–74.

6. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493: 216–220.