Applied & Translational Genomics 4 (2015) 4–9

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

NGS meta data analysis for identification of SNP and INDEL patterns in human airway transcriptome: A preliminary indicator for lung cancer

Sathya B. [a], Akila Parvathy Dharshini [b], Gopal Ramesh Kumar [b,*]

[a] Department of Bioinformatics, School of Bio Engineering, SRM University, Chennai 603203, India
[b] Department of Bioinformatics, AU KBC Research Centre, Anna University, MIT Campus, Chennai 600044, India

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 5 December 2014
Accepted 8 December 2014

Keywords:
Next generation sequencing (NGS)
Lung cancer
SNP
INDEL
Airway transcriptome
Secretoglobin

ABSTRACT

High-throughput sequencing of RNA (RNA-Seq) was developed primarily to analyze global gene expression in different tissues. It is also an efficient way to discover coding SNPs and when multiple individuals with different genetic backgrounds were used, RNA-Seq is very effective for the identification of SNPs. The objective of this study was to perform SNP and INDEL discoveries in human airway transcriptome of healthy never smokers, healthy current smokers, smokers without lung cancer and smokers with lung cancer. By preliminary comparative analysis of these four data sets, it is expected to get SNP and INDEL patterns responsible for lung cancer. A total of 85,028 SNPs and 5738 INDELs in healthy never smokers, 32,671 SNPs and 1561 INDELs in healthy current smokers, 50,205 SNPs and 3008 INDELs in smokers without lung cancer and 51,299 SNPs and 3138 INDELs in smokers with lung cancer were identified. The analysis of the SNPs and INDELs in genes that were reported earlier as differentially expressed was also performed. It has been found that a smoking person has SNPs at position 62,186,542 and 62,190,293 in SCGB1A1 gene and 180,017,251, 180,017,252, and 180,017,597 in SCGB3A1 gene and INDELs at position 35,871,168 in NFKBIA gene and 180,017,797 in SCGB3A1 gene. The SNPs identified in this study provides a resource for genetic studies in smokers and shall contribute to the development of a personalized medicine. This study is only a preliminary kind and more vigorous data analysis and wet lab validation are required.

1. Introduction

Next-generation sequencing (NGS) technology has produced immense biological data and has shed light on the path towards personalized medicine (Liu et al., 2013). NGS technology is used extensively for various applications such as: de novo sequencing, disease mapping, and quantifying expression levels through RNA sequencing etc. (Nielsen et al., 2011). The general application of NGS is: SNP and other variations discoveries, whose downstream usefulness is linkage map construction, genetic diversity analyses, association mapping and marker assisted selection. They make up over 90% of all human genetic variations, implicated in phenotype differences, risk to certain diseases and response to drugs. They also serve as popular biomarkers in pharmacogenomic studies to understand inter-individual differences in response to various treatments. Even synonymous SNPs also influence mRNA stability, protein conformation and protein regulation (Sauna and Kimchi-Sarfaty, 2011). Therefore, it is essential to obtain accurate SNP and INDEL profile information through advanced methods such as next-generation sequencing technologies (Yu and Sun, 2013). (See Fig. 1.)

The airway epithelium constitutes an essential tissue barrier protecting the lung from inhaled environmental challenges. Tobacco smoking is the dominant causative for lung/pulmonary cancer and because of this, the epithelial cells of the respiratory tract were impaired in lung cancer patients (Beane et al., 2011a; Shields, 1999; Spira et al., 2004; Miyazu et al., 2005). It creates a field of injury in epithelial cells that line the respiratory tract and is a causative factor for chronic obstructive pulmonary disease and lung cancer, with 10% to 20% of smokers developing these diseases. A smoking-related gene and miRNA expression alteration in the cytologically normal large and small airway epithelium has been proved by using microarray experiments. These expression alterations have been categorized by their degree of reversibility upon smoking cessation, providing insights into genomic changes that may account for persistent lung cancer risk. Similar gene expression alterations have been found in the epithelia of the nose and mouth of smokers (Guo et al., 2004). The serious challenge in lung cancer is that the early detection and anomalous changes in the immune system, persistent inflammation, alteration in chemokine receptor signaling and cytokine trafficking are the most essential features in disease pathogenesis (Beane et al., 2011b; Kunkel

* Corresponding author at: Bioinformatics Lab, AU-KBC Research Centre, Anna University, MIT Campus, Chrompet, Chennai 600044, India. Tel./fax: +91 44 22232711.
E-mail address: gramesh@au-kbc.org (G.R. Kumar).

http://dx.doi.org/10.1016/j.atg.2014.12.003

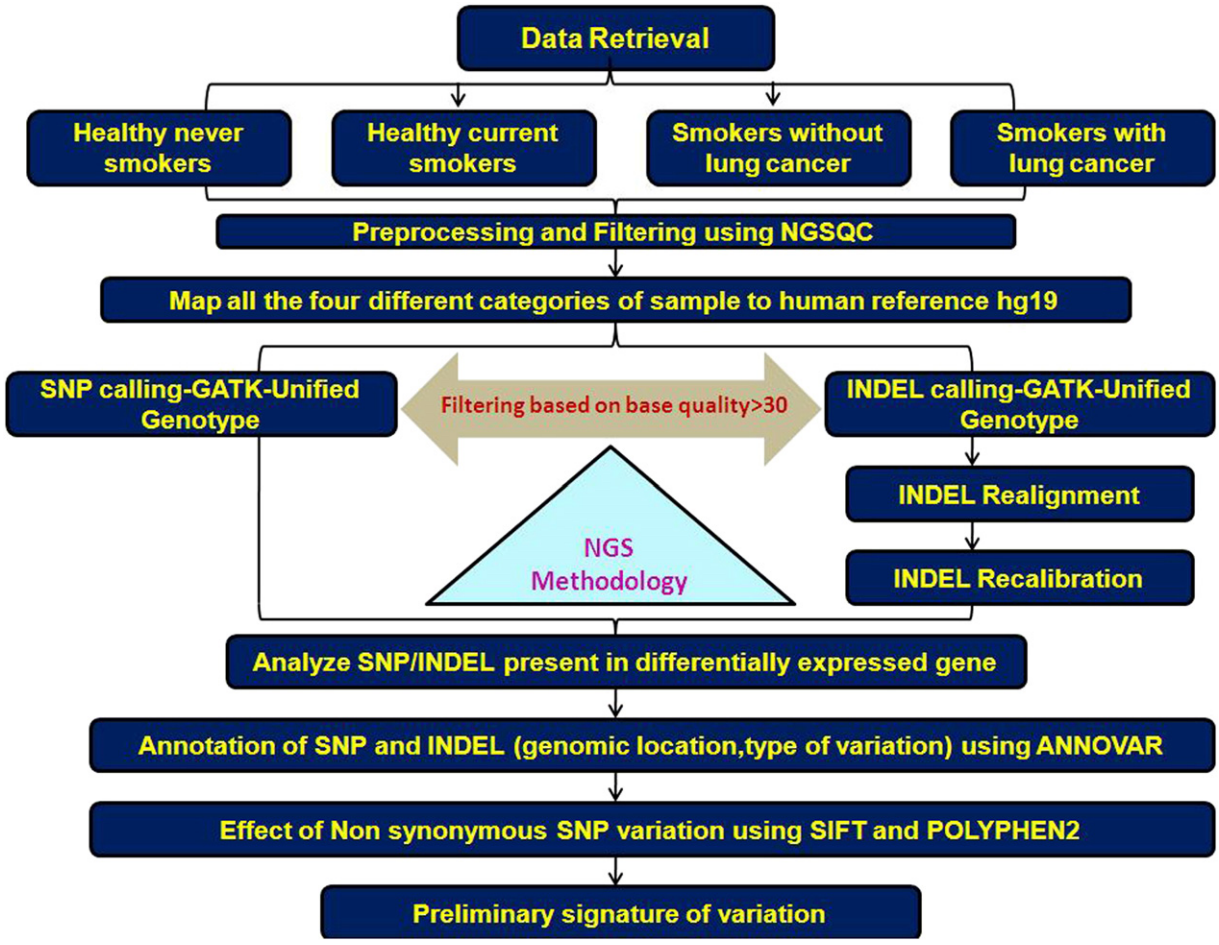

Fig. 1. Schematic representation of workflow of this current study.

and Butcher, 2002; Keith and Miller, 2013; Schembri et al., 2009; Boyle et al., 2010; Beane et al., 2007; Zhang et al., 2010; Harvey et al., 2007). Genes S100A8 and S100A9, which are known to be involved in the inflammatory response in the lung, and CYP4F2, a member of the cytochrome P450 family of enzymes that play a role in xenobiotic pathways, were found to be upregulated in smokers by both RNA-Seq and qRT-PCR. Similarly, the expression of the CCL20, IL8, NFKB1A, and SCGB3A1 immunomodulatory genes were found to be upregulated in the normal airway of patients with lung cancer (Beane et al., 2011b).

Genes involved in lung cancer from literature studies were collected and the most relevant genes in lung cancer include the following: *EGFR*, *KRAS*, *MET*, *LKB1*, *BRAF*, *PIK3CA*, *ALK*, *RET*, and *ROS1*. Other frequently mutated genes include tyrosine kinases such as EGFR homolog ERBB4 and multiple ephrin receptor genes such as EPHA3, VEGFR2 (KDR), and NTKR. Recent advances in the fields of mutational analysis and molecularly targeted therapy made it possible to develop new receptor kinase inhibitors such as erlotinib and gefitinib (against EGFR) and most recently crizotinib (against ALK) and antibodies such as ascetuximab (against EGFR) and bevacizumab (against VEGF).

In the current study, SNPs and INDELs were analyzed in four different categories of samples (healthy never smokers, current smokers, smokers with lung cancer and smoking without lung cancer) processed by RNA Seq technology, these following data procured from Frank Schembri et al., 2011 (Beane et al., 2011a). Reads were pre-processed by quality checking and filtering of low quality reads. The processed paired-end reads were mapped to reference genome hg19. Reads that mapped to reference were detached and the replica reads were removed to knock out the false SNPs and INDELs. SNP calling was implemented with default parameters and it was annotated to predict the functional impact of variants. Genes that showed distinctive expression by differential expression, gene analysis and substantiated through microarray were reported earlier in these samples (Beane et al., 2011a). The variants existing in the samples and SNPs/INDELs which are expedient to cause lung cancer in smoking person were identified. This is only a preliminary analysis and further in depth data analysis and wet lab experiments would help for further validation.

2. Materials and methods

2.1. Source of NGS data

The mRNA/transcriptome sequence of human airway epithelial cells was attained in the interim of bronchoscopy undergoing lung nodule resection surgery. These data were retrieved from prior study tabulated in Table 1 (Beane et al., 2011a). The reads were paired-end and sequenced through pooled sequence approach using Illumina technology.

The differential expression analysis was performed in these samples. The genes that are differently expressed by RNA differential expression analysis and microarray analysis between these samples are tabulated in Table 2. Pathways and molecular functions such as oxidoreductase activity, metabolism of xenobiotics by cytochrome P450 and retinol metabolism were enriched among genes differentially expressed between current and never smokers. Cytokine–cytokine receptor interaction, chemokine signaling pathway, and cell adhesion molecules were enriched among genes differentially expressed between smokers with

Table 1
Details of samples used for NGS data analysis.

Data accession	Category	Details	Age	Status
SRR192333	Healthy never smokers	3 female	~29	Healthy normal
SRR192334	Current smokers	1 male 2 female	~41.7	Smokers
SRR192335	Smokers without lung cancer	1 male 2 female	~49	2 former and 1 current smokers
SRR192336	Smokers with lung cancer	2 male 1 female	~64.7	2 former and 1 current smokers with cancer

and without cancer. The differential expressed genes may be induced by other alternations such as fusion genes, copy number variations and methylation.

2.2. Data retrieval

The next generation paired-end sequencing data were procured from the DDBJ DRA database. The data are freely accessible with the successive accession numbers SRR192333, SRR192334, SRR192335, SRR192336 for four different categories of samples such as healthy never smokers, healthy current smokers, smokers without lung cancer and smokers with lung cancer which have been analyzed in this investigation.

2.3. Pre-processing

Next generation sequencing approach generates plenty of sequence data in a single experimental run. Hence the data encompasses sequence artifacts that include read errors, poor quality reads and primer/adaptor contamination which have an impact on downstream investigation. Henceforth the quality of data is very decisive otherwise they may lead to spurious conclusions. The initial step after accomplishing the sequencing run is to assess the base quality and to trim or correct bases that do not reconcile the delineated from required specification. Pre-processing was executed using NGSQC toolkit to assess the quality of the data, examine the distribution of nucleotide, and percolate the low quality reads based on sequence constitution (Patel and Jain, 2012).

2.4. Alignment/mapping

Various groups effectively developed algorithms and software kit to execute alignment and mapping. In this work, we used Burrows–Wheeler Aligner to efficiently align short sequencing reads versus human reference genome hg19 (Pabinger et al., 2013; Li and Durbin, 2009). The accurately mapped reads were detached using the Filter SAM program in SAMtools and the duplicates were discarded using Mark duplicates in Picard tool (Li et al., 2009).

2.5. Variant calling

SNP, INDEL and structural divergent regions were precisely pinpointed using a variant calling process between four different categories of samples and the reference genome. In this study, we implemented variant calling using Genome Analysis Toolkit Unified Genotypic caller (GATK-UGT) (DePristo et al., 2011) and alignments were recalibrated around insertion/deletion region according to Freud-scaled quality scores. GATK-UGT generates output in VCF format. Along with the SNP and its position, it also reports additional information about the called SNPs, such as quality by depth (coverage), mapping quality, read depth, and genotype quality that represent the quality of the called SNPs (Ruffalo et al., 2011; Yu and Sun, 2013).

2.6. Variant annotation

It is essential to envision that the possibility of the functional impact of variants in an automated fashion is becoming progressively critical. The tools used for annotation are as follows: ANNOVAR is an efficient tool to elucidate functional residual of genetic variation and illustrated based on gene and its location and type of variation (Wang et al., 2010). SIFT was used to predict the effect of amino acid substitutions on the protein function (Ng and Henikoff, 2001). PolyPhen-2 detects the impact of structural changes intern how it is affecting the protein function by divergent sequence and homology based phylogenetic methods (Adzhubei et al., 2011).

3. Results and discussion

A simple pipeline to perceive SNPs and INDELs includes, pre-processing the sequence data and filtering low quality bases, mapping reads to the human reference genome, and post-processing of the alignment results in order to find the effect of variation.

3.1. Pre-processing

Initially the sequencing quality was scrutinized using FastQC tool. NGSQC toolkit was used to filter the low quality reads and discard the primer/adapter contaminated reads with default parameters. After filtering based on the quality score, 20.2 million reads in healthy never smokers, 10.3 million reads in healthy current smokers, 10.8 million reads in smokers without lung cancer and 10.7 million reads in smokers with lung cancer were retained and used for further analysis.

3.2. Alignment

Short sequencing reads were mapped to the annotated human reference genome (hg19) using BWA with the default parameters. Properly

Table 2
List of differentially expressed genes in diverse sample.

gene symbol	Gene name	Location	Gene expression/sample
S100A8	Calgranulin A	Chr1: 153,362,50–153,363,664	Upregulation in smokers
S100A9	Calgranulin B	Chr1: 153,330,330–153,333,503	Upregulation in smokers
CYP4F2	Cytochrome p450	Chr19: 15,988,834–6,008,885	Upregulation in smokers
NFKB1A1	NFKB inhibitor	Chr14: 35,870,717–35,873,952	Upregulation in smokers with lung cancer
SCGB1A1	Secretoglobin	Chr11: 62,172,575–62,190,667	Differentially expressed smokers with lung cancer
SCGB3A1	Secretoglobin	Chr5: 180,017,103–180,018,540	Upregulation in smokers with lung cancer
CCL20	Chemokine	Chr2: 228,678,558–228,682,272	Upregulation in smokers with lung cancer
IL8	Interleukin 8	Chr4: 74,606,223–74,609,433	Upregulation in smokers with lung cancer
RP11-295J3.2	ncRNA	Chr10: 127,660,757–27,661,695	Down regulation in smokers with lung cancer and smokers
CTD-2325P2.2	Pseudogene	Chr14: 69,159,807–69,160,300	Upregulation in smokers with lung cancer, Down regulation in smokers

Table 3
Number of SNPs present in four categories.

Category	SNP	Transition	Transversion	Ti/Tv ratio
Healthy never smokers	85,028	55,314	29,714	1.86
Healthy current smokers	32,671	21,185	11,486	1.84
Smokers without lung cancer	50,205	32,820	17,385	1.88
Smokers with lung cancer	51,299	33,063	18,236	1.81

mapped reads were separated from the unmapped reads using Filter SAM by setting the flag values in SAMtool. The main rationales for these unmapped reads are sequencing flaws, uneven quality of the sample preparations, physical gap of the reference and the defined mapping criteria. SAMtools flagstat was used to implement elementary statistics on aligning binary alignment (BAM) files. Among 44,503,612 reads, about 89.36% were aligned against hg19 and the properly mapped reads were 72.95% in healthy never smokers. Among 27,548,608 reads about 87.74% were aligned and properly mapped reads were 71.48% in healthy current smokers. Among 37,108,950 reads about 90.72% were aligned and properly mapped reads were 77.13% in smokers without lung cancer. Among 35,174,558 reads about 90.92% were aligned and properly mapped reads were 77% in smokers with lung cancer. After performing alignment, SAMtools was used to remove duplicate reads (Li et al., 2009).

Short-read alignment tools often misalign reads around INDELs, which in many cases results in mismatches. Local realignment around INDELs revamps the accuracy of INDEL calling. Local realignment, eliminates millions of mostly false positive variants while preserving nearly all truly variable sites. To attain perfect call set possible realignment and recalibration were done using INDEL REALIGNER and TABLE RECALIBRATION tools incorporated within GATK. The recalibrated alignment files were then used for SNP disclosure.

3.3. Variant calling

3.3.1. SNP/INDEL calling and annotation

The variant discovery software suite developed by the 1000 Genomes Project, the Genome Analysis Tool Kit Unified genotype caller (GATK-UGT) was used to identify SNPs and INDEL. GATK shows a relatively higher positive calling rate and sensitivity when compared to the others, and tends to call more SNPs and lower the false detection. The number of SNPs identified using GATK-UGT is tabulated in Table 3.

It was proclaimed that Ti/Tv ratio for a random variation resulting from systematic errors in the sequencing technology, alignment artifacts and data processing failures should be close to 0.5 (Ni et al., 2012; Ding et al., 2010). In the current study, transition to transversion ratio ranges from 1.81 to 1.88 that signifies that SNPs were likely resulting from true nucleotide polymorphism. SNP was sorted based on functional classes as missense, nonsense and silent mutations. Missense mutation was higher than silent and non-sense mutations. The number of missense mutations in healthy never smokers was 8571, healthy current smokers was 3005, smokers without lung cancer was 3725 and smokers with lung cancer was 3930. Further analysis with this missense mutation can identify mutated genes culpable for disease. Using ANNOVAR, annotation was performed against dbSNP to identify the known SNPs. SNPs which were not found in dbSNP are considered to be novel. The number of known and novel SNPs/INDEL was tabulated in Table 4. INDEL was called using GATK tool with the default parameters. All the QUAL scores were found to be greater than 30 and read depth approximately ranges from 20–250 which was ordinarily used as principle for calling variants in GATK. The number of insertions was found to be double than deletions in all samples. INDEL annotation was performed using an ANNOVAR tool to discover the known and novel INDELs and is tabulated in Table 4.

The annotation was performed based on genomic location and the SNPs and INDELS were distributed in exonic, intronic, intergenic, downstream, upstream, UTR3, UTR5, and splicing region. SNPs/INDELs found in the exonic region are of most influential, variation in the coding region leads to protein non-functional and it is tabulated in Table 5. INDELs found in exonic region were classified into groups as frameshift deletion and insertion, non-frameshift deletion and insertion and are tabulated in Table 5. The frameshift mutation results in abnormal protein products and changes the function and its regulation.

3.3.2. Analysis of differentially expressed genes

SNPs and INDELs were found in the following genes such as S100A8, S100A9, IL8, NFKB1A, SCGB1A1 and SCGB3A1. CYP4F2, CCL20, RP11-295J3.2, and CTD-2325P2.2 genes do not have any SNPs and INDELs. S100A8 and S100A9 genes are calcium and zinc binding proteins, which play a prominent role in regulation of inflammatory response as well as in cancer development and differential expression of these genes is associated with the disease cystic fibrosis (Ni et al., 2012; Ding et al., 2010; Ding and Kaminsky, 2003; Lim et al., 2009; Hsu et al., 2009; Henke et al., 2006). IL8 plays an essential role in pathogenesis of bronchiolitis, a common respiratory tract disease. NFKB1A is involved in inflammatory response and cancer. SCGB1A1 is a member of secretoglobin family and defects in this gene are associated with susceptibility to asthma, lung disease, respiratory failure etc. (Sjodin et al., 2003). SCGB3A1 is involved in regulation of cell proliferation and includes cytokine activity. The number of SNPs and INDELs found within these differential expressed genes is tabulated in Table 6.

S100A8, S100A9, IL8, NFKB1A and SCGB3A1 genes were upregulated in the disease pathogenesis. Based on the category of sample the common variation which is present in smokers with lung cancer, smokers without lung cancer and current smokers was tabulated in Table. 7.

Table 4
Number of Known and Novel SNPs/INDEL present in all the four categories.

Category	Total SNPs	Known SNPs	Novel SNPs	Total INDELs	Known INDELs	Novel INDELs
Healthy never smokers	85,028	37,635	47,393	5738	2305	3433
Healthy current smokers	32,671	14,396	18,275	1561	654	910
Smokers without lung cancer	50,205	21,914	28,291	3008	1300	1708
Smokers with lung cancer	51,299	21,451	29,848	3138	1363	1775

Table 5
SNPs/INDEL present in exonic/functional region.

Category	Synonymous-SNP	Non-synonymous-SNP	Frameshift deletion	Frameshift insertion	Non-frameshift deletion	Non-frameshift insertion
Healthy never smokers	3116	3234	71	230	10	24
Healthy current smokers	1419	1429	23	71	4	9
Smokers without lung cancer	1830	2061	46	107	4	19
Smokers with lung cancer	1902	2058	48	116	5	13

Table 6
Number of SNPs and INDELs present in differentially expressed genes.

Category	SNPs	INDELs
Healthy never smokers	39	14
Healthy current smokers	29	3
Smokers without lung cancer	27	7
Smokers with lung cancer	43	10

These variations were not observed in healthy never smokers and the following changes detected in regulatory region that may affect the regulation of the gene.

When evaluating the changes between smokers with lung cancer and without lung cancer, we detect some discrepancy in the IL8 regulatory region. By comparing the entire sample, some of the variations are present only in smokers with lung cancer.

The effect of non-synonymous aberrations can be anticipated using SIFT and PolyPhen-2. SIFT predicts the potential impact of amino acid substitution on protein function based on sequence homology. PolyPhen-2 also detects the effect of non-synonymous variation based on homologous sequence and structure based prediction. SNP present in this position 180,017,251, amino acid changes alanine to serine predicted to be deleterious based on Polyphen-2 prediction. Frameshift insertion in the position 180,017,797 envisioned to be damaging the protein function. The above SNPs were not reported in catalogue of somatic mutation in cancer (COSMIC) database (Forbes et al., 2011).

4. Conclusion

Transcriptome analysis using next generation sequencing is the most competent and cost effective for identification of SNPs and INDELs. In the current study, our main focus was to analyze the variants existing in the genes which were disclosed earlier as differentially expressed. We found 5 SNPs and 2 INDELs in SCGB1A1, SCGB3A1 and NFKB1A genes which were present only in smokers with lung cancer. Hence a smoking person having this set of SNPs and INDELs is a preliminary signature for the disease pathogenesis. Understanding an individual's genetic makeup is believed to be the key role in personalized medicine to maximize drug efficacy and minimize adverse side effects. In the future, an individual's genome will be sequenced to predict the future health of the individual and to develop personalized medical treatments that are tailored to work with the genetic variation that is detected (Mullaney et al., 2010). Extending this study to a large group of population will be useful for developing personalized medicine. In the future, we are planning to study the stability of wild and mutant protein for secretoglobin using extensive molecular dynamics study in order to find the effect of frameshift and non-synonymous mutation in structural level.

Acknowledgment

The authors are thankful to PhD Student Nupoor Chowdhary for her support and technical assistance.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2011. A method and server for predicting damaging missense mutations. Nat. Methods 7 (4), 248–249.

Beane, J., Sebastiani, P., Liu, G., Brody, J.S., Lenburg, M.E., Spira, A., 2007. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol. 8 (9), R201.

Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X.H., Xiao, J., Alekseyev, Y.O., Wang, S., Levy, S., Massion, P.P., Lenburg, M., Spira, A., 2011a. Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. Cancer Prev. Res. (Phila.) 4 (6), 803–817.

Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X.H., Xiao, J., Alekseyev, Y.O., Wang, S., Levy, S., Massion, P.P., Lenburg, M., Spira, A., 2011b. Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. Cancer Prev. Res. 4 (6), 803–817.

Boyle, J.O., Gümüs, Z.H., Kacker, A., Choksi, V.L., Bocker, J.M., Zhou, X.K., Yantiss, R.K., Hughes, D.B., Du, B., Judson, B.L., Subbaramaiah, K., Dannenberg, A.J., 2010. Effects of cigarette smoke on the human oral mucosal transcriptome. Cancer Prev. Res. 3, 266–278.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43 (5), 491–498.

Ding, X., Kaminsky, L.S., 2003. Human extrahepatic cytochromes P450: function in xenobiotic metabolism and tissue-selective chemical toxicity in the respiratory and gastrointestinal tracts. Annu. Rev. Pharmacol. Toxicol. 43, 149–173.

Ding, L., Wendl, M.C., Koboldt, D.C., Mardis, E.R., 2010. Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum. Mol. Genet. 19 (R2), R188–R196.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., Futreal, P.A., 2011. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39 (Database issue), D945–D950.

Guo, M., House, M.G., Hooker, C., Han, Y., Heath, E., Gabrielson, E., Yang, S.C., Baylin, S.B., Herman, J.G., Brock, M.V., 2004. Promoter hypermethylation of resected bronchial margins: a field defect of changes? Clin. Cancer Res. 10, 5131–5136.

Harvey, B.G., Heguy, A., Leopold, P.L., Carolan, B.J., Ferris, B., Crystal, R.G., 2007. Modification of gene expression of the small airway epithelium in response to cigarette smoking. J. Mol. Med. 85, 39–53.

Henke, M.O., Renner, A., Rubin, B.K., Gyves, J.I., Lorenz, E., Koo, J.S., 2006. Up-regulation of S100A8 and S100A9 protein in bronchial epithelial cells by lipopolysaccharide. Exp. Lung Res. 32, 331–347.

Hsu, K., Champaiboon, C., Guenther, B.D., Sorenson, B.S., Khammanivong, A., Ross, K.F., Geczy, C.L., Herzberg, M.C., 2009. Antiinfective protective properties of S100 calgranulins. Antiinflamm. Antiallergy Agents Med. Chem. 8 (4), 290–305.

Keith, R.L., Miller, Y.E., 2013. Lung cancer chemoprevention: current status and future prospects. Nat. Rev. Clin. Oncol. 10, 334–343.

Table 7
SNPs and INDELs present in different categories.

Chr	Position	dbSNP	Ref/Alt	Category	Location	Gene	Type
1	153,333,376	Novel	A/C	CS + SNL + SL	UTR3	S100A9	NA
4	74,606,669	rs2227307	T/G	CS + SNL + SL	Intronic	IL8	NA
1	153,362,719	Novel	T/TC	CS + SNL + SL	Intronic	S100A8	Frameshift insertion
4	74,607,910	rs2227543	C/T	SNL + SL	Intronic	IL8	NA
4	74,608,162	Novel	T/A	SNL + SL	Intronic	IL8	NA
4	74,608,163	Novel	T/G	SNL + SL	Intronic	IL8	NA
4	74,608,408	Novel	C/CA	SNL + SL	UTR3	IL8	NA
11	62,186,542	rs3741240	G/A	SL	UTR5	SCGB1A1	NA
11	62,190,293	rs191704193	G/A	SL	Intronic	SCGB1A1	NA
5	180,017,251	Novel	C/A	SL	Exonic	SCGB3A1	Nonsynonymous SNV
5	180,017,252	Novel	C/G	SL	Exonic	SCGB3A1	Synonymous SNV
5	180,017,597	Novel	T/C	SL	Intronic	SCGB3A1	NA
14	35,871,168	Novel	C/CT	SL	UTR3	NFKB1A	NA
5	180,017,797	Novel	G/GA	SL	Exonic	SCGB3A1	Frameshift insertion

NA: not Available, CS: current smokers, SNL: smokers with no lung cancer, SL: smokers with lung cancer.

Kunkel, E.J., Butcher, E.C., 2002. Chemokines and the tissue-specific migration of lymphocyte. Immunity 16 (1), 1–4.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (14), 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25 (16), 2078–2079.
Lim, S.Y., Raftery, M.J., Goyette, J., Hsu, K., Geczy, C.L., 2009. Oxidative modifications of S100 proteins: functional regulation by redox. J. Leukoc. Biol. 86 (3), 577–587.
Liu, X., Han, S., Wang, Z., Gelernter, J., Yang, B.Z., 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS ONE 8 (9), e75619 (September 27).
Miyazu, Y.M., Miyazawa, T., Hiyama, K., Kurimoto, N., Iwamoto, Y., Matsuura, H., Kanoh, K., Kohno, N., Nishiyama, M., Hiyama, E., 2005. Telomerase expression in noncancerous bronchial epithelia is a possible marker of early development of lung cancer. Cancer Res. 65, 9623–9627.
Mullaney, J.M., Mills, R.E., Pittard, W.S., Devine, S.E., 2010. Small insertions and deletions (INDELs) in human genomes. Hum. Mol. Genet. 19 (R2), R131–R136.
Ng, P.C., Henikoff, S., 2001. Predicting deleterious amino acid substitutions. Genome Res. 11 (5), 863–874.
Ni, Y., Hall, A.W., Battenhouse, A., Iyer, V.R., 2012. Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. BMC Genet. 13, 46.
Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12 (6), 443–451.
Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z., 2013. A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. 15 (2), 256–278.
Patel, R.K., Jain, M., 2012. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE 7 (2), e30619.
Ruffalo, M., LaFramboise, T., Koyutürk, M., 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27 (20), 2790–2796.
Sauna, Z.E., Kimchi-Sarfaty, C., 2011. Understanding the contribution of synonymous mutations to human disease. Nat. Rev. Genet. 12 (10), 683–691.
Schembri, F., Sridhar, S., Perdomo, C., Gustafson, A.M., Zhang, X., Ergun, A., Lu, J., Liu, G., Zhang, X., Bowers, J., Vaziri, C., Ott, K., Sensinger, K., Collins, J.J., Brody, J.S., Getts, R., Lenburg, M.E., Spira, A., 2009. MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. Proc. Natl. Acad. Sci. U. S. A. 106, 2319–2324.
Shields, P.G., 1999. Molecular epidemiology of lung cancer. Ann. Oncol. 10 (Suppl. 5), S7–S11.
Sjodin, A., Guo, D., Sorhaug, S., Bjermer, L., Henriksson, R., Hedman, H., 2003. Dysregulated secretoglobin expression in human lung cancers. Lung Cancer 41 (1), 49–56.
Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., Brody, J.S., 2004. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc. Natl. Acad. Sci. 101, 10143–10148.
Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38 (16), e164.
Yu, X., Sun, S., 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinforma. 14, 274.
Zhang, X., Sebastiani, P., Liu, G., Schembri, F., Zhang, X., Dumas, Y.M., Langer, E.M., Alekseyev, Y., O'Connor, G.T., Brooks, D.R., Lenburg, M.E., Spira, A., 2010. Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. Physiol. Genomics 41 (1), 1–8.